Chapter 12

Fermat's Last Theorem and unique factorisation

Introduction

In the last chapter, we began our exploration of ring theory by looking at the specific case of polynomial rings over fields. We learnt how to carry out polynomial division over arbitrary fields and compute highest common factors, and we then established conditions for irreducibility of polynomials.

In this chapter, we will extend our use of the word 'irreducible' to arbitrary rings and use this to explore how we may generalise the *Fundamental Theorem of Arithmetic*; that is, we will ask when we can factorise an element of a ring R into irreducible elements, and when this can be done uniquely. As you will see, this story of unique factorisation relates strongly to the solutions of certain Diophantine equations and it is with the story of Fermat's Last Theorem that our investigation begins.

Do not confuse this with the Fundamental Theorem of Algebra, which concerns irreducible polynomials over $\mathbb{C}[x]$.

Chapter guide

This chapter has four sections. The first section, although setting the theme for the rest of the chapter, does not contain mathematics used by the later sections. The three later sections, however, do build on each other, and they contain material that will be used in Book E. These three sections are of roughly equal length, although the second section contains fewer technical arguments than the final two.

1 Fermat's Last Theorem and Diophantine equations

Fermat's refusal to publish his discoveries led to one of the great stories in the history of mathematics. As well as writing letters to his friends about various results, he formed the habit of jotting notes in the margins of his reference books. Some five years after his death, his copy of Bachet's *Diophantus* was found to contain mention of many results from number theory. On the page dealing with the Pythagorean equation, Fermat had added the following in the margin.

> On the other hand it is impossible to write a cube as the sum of two cubes, a fourth power as the sum of two fourth powers and, in general, any power higher than the second as the sum of two similar powers. I have discovered a truly marvellous proof of this, but the margin is too small to contain it.

Here, Fermat was claiming to have proved that the Diophantine equation $x^n + y^n = z^n$ has no positive integer solutions for $n \geq 3$. No evidence of

how Fermat reached his conclusions survives, and it took another three and a half centuries of continual efforts to solve the problem. As this was the last of Fermat's results awaiting proof or refutation, it became known colloquially as *Fermat's Last Theorem*. Strictly speaking, until its proof it should not have been called a theorem, and you will find it referred to as Fermat's Conjecture in some older texts.

Consider the equation $x^n + y^n = z^n$. When $n = 2$, this is the Pythagorean equation, and we will classify the integer solutions to this in the next subsection. For $n \geq 3$, if the exponent n is composite, say $n = kr$, then the equation can be rewritten as

$$(x^k)^r + (y^k)^r = (z^k)^r.$$

It follows that if $x = x_0$, $y = y_0$, $z = z_0$ is a solution of $x^n + y^n = z^n$ then

$$x = x_0^k, \quad y = y_0^k, \quad z = z_0^k$$

is a solution of $x^r + y^r = z^r$. Conversely, if it can be established that $x^r + y^r = z^r$ has no positive solutions then it follows that $x^n + y^n = z^n$ must have no positive solutions.

Now, every integer $n \geq 3$ is divisible by some odd prime or by 4. The proof of Fermat's Last Theorem can therefore be reduced to proving two cases:

1. $x^4 + y^4 = z^4$ has no positive solution
2. $x^p + y^p = z^p$ has no positive solution for any odd prime p.

In this section, after we have solved the Pythagorean equation, we will present the method of 'infinite descent' and use this, as Fermat himself did, to prove the case $n = 4$. The real challenge, however, lay in proving Fermat's Last Theorem for odd prime exponents, and we will pick up this story at the start of the next section.

1.1 Pythagorean triples

Pythagoras' Theorem, which states that the sum of the squares of the two shorter sides of a right-angled triangle is equal to the square of the hypotenuse (Figure 1.1), is arguably the most celebrated theorem in mathematics. But in addition to being a classical theorem of geometry, it presents challenges for the number theorist when the underlying equation $x^2 + y^2 = z^2$ is considered as a Diophantine equation. We refer to this as the **Pythagorean equation**.

We will continue the geometric analogy by referring to the value of z in any solution as being the **hypotenuse**.

Of course, unlike the equation in Fermat's Last Theorem, the Pythagorean equation $x^2 + y^2 = z^2$ has non-trivial solutions in the natural numbers, including the well-known one that has the smallest value of z, namely $3^2 + 4^2 = 5^2$. It is also immediate from this one solution that the equation has infinitely many solutions since, for any positive integer k,

$$(3k)^2 + (4k)^2 = (5k)^2.$$

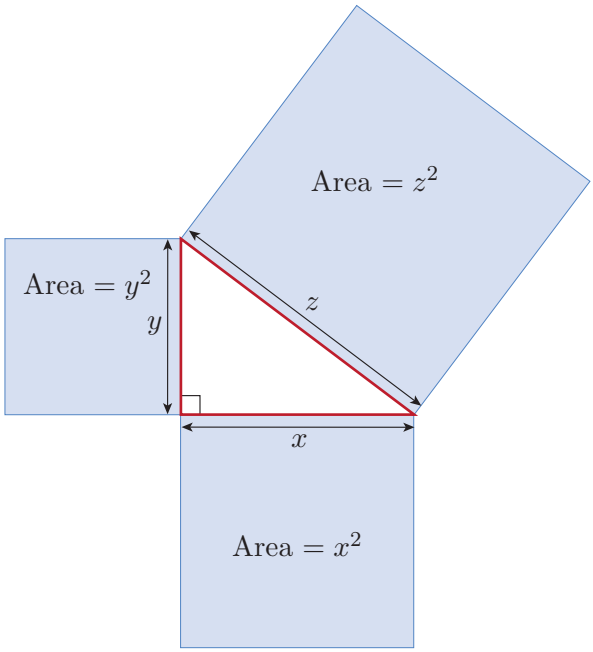


Figure 1.1 Pythagoras' Theorem: $x^2 + y^2 = z^2$

But there is a sense in which this infinite family of solutions is really just one solution. Geometrically, the corresponding triangles are similar; we have the one basic triangle with sides 3, 4 and 5, and the others are obtained by scaling each side by factor k.

Before progressing further, let us introduce some terminology.

Definition 1.1 *Pythagorean triples*

A **Pythagorean triple** is a triple (x, y, z) of positive integers such that $x^2 + y^2 = z^2$.

The triple is said to be **primitive** if $\text{hcf}(x, y) = 1$.

A triple is ordered so that the hypotenuse is always the third member.

Notice that the condition $\text{hcf}(x, y) = 1$ guarantees that x and y have no common factor greater than 1, and consequently the primitive triple (x, y, z) cannot be a scaling of some smaller triple. Moreover, the condition $\text{hcf}(x, y) = 1$ carries the implications that $\text{hcf}(x, z) = \text{hcf}(y, z) = 1$, since from the equation $x^2 + y^2 = z^2$ it is readily shown that any common divisor of two of the variables must divide the third.

Exercise 1.1

Which, if any, of the following are Pythagorean triples? For each Pythagorean triple, decide whether or not it is primitive.

(a) $(10, 8, 6)$

(b) $(12, 9, 15)$

(c) $(6, 7, 8)$

(d) $(5, 12, 13)$

(e) $(24, 33, 41)$

We now have two primitive Pythagorean triples, namely $(3, 4, 5)$ and $(5, 12, 13)$. Are there any more? The Babylonians, some 3500 years ago, were aware of many primitive solutions, but Heath, in Book I of his translation of Euclid's *Elements*, credits Pythagoras with the first infinite family of solutions given by

$$x = 2k + 1, \quad y = 2k^2 + 2k, \quad z = 2k^2 + 2k + 1, \quad \text{for any integer } k \geq 1.$$

The first complete solution of the Pythagorean equation appeared in Euclid's *Elements*. We aim to reproduce that solution here, but first there are a few points to be clarified.

If (x, y, z) is any Pythagorean triple then x and y cannot both be odd. For if they were then

$$z^2 = x^2 + y^2 \equiv 1 + 1 \equiv 2 \,(\text{mod } 4),$$

which is impossible as all squares are congruent modulo 4 to either 0 or 1. It follows that in a primitive triple x and y have opposite parity, because the further condition that $\text{hcf}(x, y) = 1$ ensures that they cannot both be even.

The **parity** of a pair of integers is defined as follows. If both are odd or both are even, they are said to have the *same* parity; otherwise they are said to have *opposite* parity.

If (x, y, z) is a primitive Pythagorean triple then so too is (y, x, z), because the x and y values can certainly be interchanged. But, to all intents and purposes, the triples (x, y, z) and (y, x, z) lead to what is really the same solution of the Pythagorean equation. To overcome the need to distinguish between these two equivalent solutions, we choose to write the even member first in any primitive triple. We adopt the following convention.

Convention for Pythagorean triples

If (x, y, z) is any primitive Pythagorean triple, then x is even while y and z are both odd.

With the preparation now complete, we are ready for the main theorem of this subsection.

Theorem 1.2 *Primitive solutions of the Pythagorean equation*

The primitive Pythagorean triples are the triples

$$(2mn, m^2 - n^2, m^2 + n^2),$$

where m and n are coprime natural numbers of opposite parity with $m > n$.

Before embarking on the proof, let us clarify the claim of the theorem by looking at an example of its use.

Worked Exercise 1.3

Find all primitive Pythagorean triples $(60, y, z)$.

Solution

We seek coprime integers m and n such that $2mn = 60$, m and n have opposite parity and $m > n$. There are four pairs of positive integers satisfying $mn = 30$ with $m > n$, and in all four cases m and n are coprime and have opposite parity:

$m = 30,\ n = 1,$ giving $(60, 899, 901)$;
$m = 15,\ n = 2,$ giving $(60, 221, 229)$;
$m = 10,\ n = 3,$ giving $(60, 91, 109)$;
$m = \ 6,\ n = 5,$ giving $(60, 11, 61)$.

These four are the only primitive Pythagorean triples $(60, y, z)$.

Proof of Theorem 1.2 We must first show that the given triple is a primitive Pythagorean triple. Substituting $x = 2mn$, $y = m^2 - n^2$ and $z = m^2 + n^2$ into the Pythagorean equation, we have

$$\begin{aligned} x^2 + y^2 &= (2mn)^2 + (m^2 - n^2)^2 \\ &= m^4 + 2m^2n^2 + n^4 = (m^2 + n^2)^2 = z^2. \end{aligned}$$

So this is indeed a Pythagorean triple. To establish the primitive property, suppose to the contrary that p is a prime factor of both $m^2 - n^2$ and $m^2 + n^2$. Then p divides both the sum $(m^2 + n^2) + (m^2 - n^2) = 2m^2$ and the difference $(m^2 + n^2) - (m^2 - n^2) = 2n^2$. But $p \neq 2$ (since y is odd) and so we conclude that p divides m and p divides n. As m and n are given to be coprime we have the required contradiction, and so the triple is primitive.

Conversely, we have to show that every primitive Pythagorean triple is of the stated form. With this goal in mind, suppose that (x, y, z) is *any* primitive Pythagorean triple. As y and z are both odd, we can define integers

$$s = \frac{z+y}{2} \text{ and } t = \frac{z-y}{2}.$$

We note that s and t are coprime because any common factor would also divide $s + t = z$ and $s - t = y$.

Then from the Pythagorean equation

$$x^2 = z^2 - y^2 = (z+y)(z-y) = 2s \times 2t$$

we obtain

$$\left(\frac{x}{2}\right)^2 = st.$$

As x is known to be even, this equation confirms that the product st is a perfect square. Therefore, since s and t are coprime, they must each be squares. So if we now write $s = m^2$ and $t = n^2$ and substitute back, we get

This follows from the work of Book A, Chapter 2, but see also Exercise 1.2(c) at the end of this subsection.

$$x^2 = 4st = 4m^2n^2, \quad \text{giving } x = 2mn;$$
$$y = s - t = m^2 - n^2;$$
$$z = s + t = m^2 + n^2.$$

Finally, note that $\text{hcf}(m, n) = 1$, because any common factor of m and n would necessarily be a common factor of the coprime integers $s = m^2$ and $t = n^2$. Moreover, m and n must have opposite parity for otherwise y and z would be even.

The triple (x, y, z) has thus been expressed in the required form. ■

All Pythagorean triples can be obtained by scaling primitive ones; that is, any Pythagorean triple is of the form (kx, ky, kz), where (x, y, z) is a primitive triple. Notice that our convention for primitive triples carries over to all triples, and so the first member of the triple is always the multiple of the even member in the underlying primitive triple. With this clarified, we can now record the following corollary.

Corollary 1.4 *All solutions of the Pythagorean equation*

The Pythagorean triples (x, y, z) are given by

$$x = 2kmn, \quad y = k(m^2 - n^2), \quad z = k(m^2 + n^2),$$

where $k \geq 1$ is any integer and m and n are coprime natural numbers with opposite parity and $m > n$.

The fact that m, n are coprime implies that there are no Pythagorean triples (x, y, z) with $x = y$.

In the next worked exercise we make use of Theorem 1.2 to begin an enumeration of all the primitive Pythagorean triples.

Worked Exercise 1.5

There are 16 primitive Pythagorean triples (x, y, z) with hypotenuse $z < 100$. Find them.

Solution

We find primitive triples by listing pairs of coprime integers m and n that have opposite parity and with $m > n$. For each such pair, the corresponding triple is calculated from the formulas in Theorem 1.2. For example, taking $m = 7$, each of the values $n = 2$, 4 and 6 will lead to a primitive triple.

For $m = 7$ and $n = 1$, 3 or 5 the resulting triple will be Pythagorean but will not be primitive, because m and n have the same parity.

The condition that $z < 100$ amounts to $m^2 + n^2 < 100$. Hence $m \leq 9$ and, for example, the triple resulting from $m = 8$, $n = 7$ is not wanted because its hypotenuse exceeds 100. The 16 solutions are as shown in Table 1.1.

Table 1.1 Primitive triples with sides not exceeding 100

m	n	triple	m	n	triple
2	1	$(4, 3, 5)$	7	2	$(28, 45, 53)$
3	2	$(12, 5, 13)$	7	4	$(56, 33, 65)$
4	1	$(8, 15, 17)$	7	6	$(84, 13, 85)$
4	3	$(24, 7, 25)$	8	1	$(16, 63, 65)$
5	2	$(20, 21, 29)$	8	3	$(48, 55, 73)$
5	4	$(40, 9, 41)$	8	5	$(80, 39, 89)$
6	1	$(12, 35, 37)$	9	2	$(36, 77, 85)$
6	5	$(60, 11, 61)$	9	4	$(72, 65, 97)$

Exercise 1.2

(a) Find all primitive Pythagorean triples $(72, y, z)$.

(b) Find all Pythagorean triples, primitive or not, in which one of the sides is 30.

(c) Give a proof of the step assumed in Theorem 1.2 that if a product of two coprime integers is a square, then each of the integers must be square.

1.2 The equation $x^4 + y^4 = z^4$

In the remainder of this section we will confine our attention to the special case of Fermat's Last Theorem for exponent 4 and some related Diophantine equations. We will use Fermat's method and, just as he did, prove the stronger result that $x^4 + y^4 = z^2$ has no positive solution. The proof uses our solution to the Pythagorean equation together with a powerful technique known as Fermat's **method of infinite descent**.

The idea behind the method of infinite descent as applied to this problem is as follows. We establish that from any positive solution $x = x_k$, $y = y_k$, $z = z_k$ of the Diophantine equation $x^4 + y^4 = z^2$ we can always find another 'smaller' positive solution $x = x_{k+1}$, $y = y_{k+1}$, $z = z_{k+1}$, where by smaller we mean that $0 < z_{k+1} < z_k$. This process of being able to go from one positive solution to a smaller one is the key stage of the method, and is called the **descent step**.

Now suppose that the equation does have a positive solution $x = x_1$, $y = y_1$, $z = z_1$. The established descent step would then tell us that, from this, we have a smaller positive solution, and from this second solution a still smaller third positive solution, from which we get a smaller fourth positive solution, and so on. Focusing on the z-values, we have an unending decreasing sequence of positive integers

$$z_1 > z_2 > z_3 > z_4 > \cdots > 0.$$

As there are only finitely many positive integers that are less than z_1, this is plainly impossible. Hence the one assumption that we have made, namely that one positive solution exists, is contradicted.

Infinite descent bears some resemblance to mathematical induction. The descent step, like the induction step, sets up an unending chain of implications of the form 'if one is true then so too is the next'. The difference is that this time we know that no such infinite chain of true statements can exist and we are forced to conclude that the chain cannot be initiated; it has no basis.

We have chosen to focus on the variable z here. But notice that descent could be achieved on any of the variables. For example, if we established that any positive solution of the equation must give rise to another positive solution with smaller value for y, then the conclusion that no solution can exist follows in the same way.

What is really being brought into play here is the Well-Ordering Principle from Book A, Chapter 1, which asserts that any non-empty set of positive integers must have a least member. The descent step appears to construct a set of positive integers that does not have a least member, and so it has to be the empty set.

Theorem 1.6

The Diophantine equation $x^4 + y^4 = z^2$ has no positive solutions.

The proof of Theorem 1.6 and that of Theorem 1.9, which follows shortly, involve some tricky algebra. You are not necessarily expected to master the details of these proofs, but you should read each of them carefully to see how the method of infinite descent is used.

Proof Suppose to the contrary that $x^4 + y^4 = z^2$ has a positive solution $x = x_1$, $y = y_1$, $z = z_1$. If $\mathrm{hcf}(x_1, y_1) = d > 1$ then putting $x_1 = dx'$ and $y_1 = dy'$ gives $z_1^2 = d^4(x'^4 + y'^4)$, from which it follows that d^2 divides z_1 so that $z_1 = d^2 z'$, for some integer $z' < z_1$. But then $x'^4 + y'^4 = z'^2$, where $\mathrm{hcf}(x', y') = 1$ and $z' < z_1$. This argument shows that we may assume that $\mathrm{hcf}(x_1, y_1) = 1$, for otherwise any common divisor can first be cancelled leaving another, smaller solution to the same equation.

Rather than use the general suffix k as in the preamble, we establish the descent step here by showing that any positive solution x_1, y_1, z_1 must lead to a positive solution x_2, y_2, z_2 in which $0 < z_2 < z_1$.

Writing the equation as $(x_1^2)^2 + (y_1^2)^2 = z_1^2$, we observe that (x_1^2, y_1^2, z_1) is a primitive Pythagorean triple and so, by Theorem 1.2,

If needed we may interchange x_1 and y_1 to ensure that x_1 is even.

$$x_1^2 = 2mn, \quad y_1^2 = m^2 - n^2, \quad z_1 = m^2 + n^2,$$

where m and n are coprime positive integers of opposite parity with $m > n$. In fact n must be even and m odd, for otherwise we have

$$y_1^2 = m^2 - n^2 \equiv 0 - 1 \equiv 3 \,(\mathrm{mod}\ 4),$$

which is impossible because any square is congruent to either 0 or 1 modulo 4.

Putting $n = 2r$, the equation for x_1 becomes

$$\left(\frac{x_1}{2}\right)^2 = mr,$$

where m and $r = \frac{n}{2}$ are coprime. Hence m and r are each squares, say $m = s^2$ and $r = t^2$.

Returning to the equation $y_1^2 = m^2 - n^2$, we see that (n, y_1, m) is a primitive Pythagorean triple and so

$$n = 2uv, \quad y_1 = u^2 - v^2, \quad m = u^2 + v^2,$$

where u and v are coprime positive integers of opposite parity with $u > v$.

Now $n = 2r = 2t^2$, and so the first of these equations becomes $uv = t^2$. It follows that u and v are each squares, say $u = x_2^2$ and $v = y_2^2$. Feeding these facts into the equation $m = u^2 + v^2$ gives

$$x_2^4 + y_2^4 = m = s^2 (= z_2^2, \text{ say}),$$

revealing another solution of the original equation. But this solution is 'smaller' since

$$0 < z_2 = s \leq m < m^2 + n^2 = z_1.$$

This completes the descent step; any positive solution gives rise to a smaller positive one. As this is impossible, the assumption that there exists a positive solution is contradicted. ■

As any fourth power is necessarily a square, any positive solution of $x^4 + y^4 = z^4$ would contradict the result of Theorem 1.6. We deduce the following.

Corollary 1.7

The Diophantine equation $x^4 + y^4 = z^4$ has no positive solutions.

The method of infinite descent used in the proof of Theorem 1.6 can be applied to many Diophantine equations, almost invariably with the purpose of showing that the equation has no positive solutions. Here is a simpler example.

Worked Exercise 1.8

Show that the Diophantine equation $x^3 + 3y^3 = 9z^3$ has no positive solutions.

Solution

Aiming for a contradiction, suppose that $x_1^3 + 3y_1^3 = 9z_1^3$, where x_1, y_1 and z_1 are positive integers. As the prime 3 divides the right-hand side and one of the terms on the left of this equation, it must divide the remaining term. That is, 3 divides x_1^3 and so 3 divides x_1. This means we can find a positive integer x_2 so that $x_1 = 3x_2$, and this gives us

$$27x_2^3 + 3y_1^3 = 9z_1^3.$$

We can divide this equation by 3, to obtain

$$9x_2^3 + y_1^3 = 3z_1^3.$$

Repetition of the same reasoning now shows that 3 divides y_1, and so there is a positive integer y_2 with $y_1 = 3y_2$. This gives

$$9x_2^3 + 27y_2^3 = 3z_1^3; \quad \text{that is, } 3x_2^3 + 9y_2^3 = z_1^3.$$

This time we have 3 dividing z_1, and putting $z_1 = 3z_2$ for some positive integer z_2, the equation now becomes

$$3x_2^3 + 9y_2^3 = 27z_2^3; \quad \text{that is, } x_2^3 + 3y_2^3 = 9z_2^3.$$

At this point we have reached a second positive solution, $x = x_2$, $y = y_2$, $z = z_2$ of the original equation with $z_2 < z_1$. The descent step is therefore complete and the required contradiction is established. Thus $x^3 + 3y^3 = 9z^3$ has no positive solutions.

Exercise 1.3

Show that the Diophantine equation $x^4 + 2y^4 = 4z^4$ has no positive solutions.

The following exercise requires a slight variant on Fermat's method of infinite descent.

Exercise 1.4

(a) Show that if $x^2 + y^2 + z^2 \equiv 0 \pmod 4$ then each of x, y and z is even.

(b) Show that in any positive solution $x = x_1$, $y = y_1$, $z = z_1$ of the Diophantine equation $x^2 + y^2 + z^2 = 2xyz$, each of x_1, y_1 and z_1 must be even. Putting $x_1 = 2x_2$, $y_1 = 2y_2$ and $z_1 = 2z_2$, show that x_2, y_2 and z_2 are also even positive integers.

(c) Prove that the Diophantine equation $x^2 + y^2 + z^2 = 2xyz$ has no positive solution by constructing from such a solution a strictly decreasing infinite sequence of even positive integers.

1.3 Related Diophantine equations

Fermat went on to apply his method of infinite descent to prove that the related equation $x^4 - y^4 = z^2$ has no positive solution. We will, however, prove this result as a corollary of the following theorem.

Theorem 1.9

The Diophantine equation $x^4 + 4y^4 = z^2$ has no positive solution.

Proof Suppose that x_1, y_1 and z_1 are positive integers with $x_1^4 + 4y_1^4 = z_1^2$, that is, $(x, y, z) = (x_1, y_1, z_1)$ is a particular solution. Exactly as in the proof of Theorem 1.6 we may assume that $\text{hcf}(x_1, y_1) = 1$, for otherwise cancellation of any common divisor leads to a smaller solution of the same equation.

With infinite descent in mind, we wish to show that the existence of this one positive solution necessarily gives rise to one with a smaller z-value. We observe that x_1 and z_1 are of the same parity, and we first consider the case in which they are both even. Substituting $x_1 = 2x'$ and $z_1 = 2z'$ gives

$$(2x')^4 + 4y_1^4 = (2z')^2; \quad \text{that is, } y_1^4 + 4x'^4 = z'^2,$$

which is another positive solution of the equation with $z' < z_1$.

The solution is $(x, y, z) = (y_1, x', z')$.

It remains to show that a solution in which x_1 and z_1 are both odd must likewise lead to another with smaller z-value. Writing the equation as $(x_1^2)^2 + (2y_1^2)^2 = z_1^2$, we recognise that $(2y_1^2, x_1^2, z_1)$ is a primitive

Pythagorean triple and so, according to Theorem 1.2, there are coprime positive integers m and n of opposite parity and with $m > n$ such that

$$2y_1^2 = 2mn, \quad x_1^2 = m^2 - n^2, \quad z_1 = m^2 + n^2.$$

From $x_1^2 = m^2 - n^2$ we see that m is odd and n is even, for the other way round would give $m^2 - n^2 \equiv 3 \pmod 4$ and this cannot be a square. And from $y_1^2 = mn$, the coprime integers m and n are each squares and so we can write $m = a^2$ and $n = (2b)^2$.

The triple is primitive because $\text{hcf}(m, n) = 1$.

Looking again at $x_1^2 = m^2 - n^2$ with $\text{hcf}(m, n) = 1$, we now see that (n, x_1, m) is another primitive Pythagorean triple, and so there exist coprime positive integers s and t, of opposite parity and with $s > t$ such that

$$n = 2st, \quad x_1 = s^2 - t^2, \quad m = s^2 + t^2.$$

Substituting $n = (2b)^2$ into the first of these equations gives $2b^2 = st$. Now one of s or t is even. If it is s, say $s = 2r$, then $b^2 = rt$ and as r and t are coprime each is a square; say $r = u^2$ (so that $s = 2u^2$) and $t = v^2$. The alternative that t is even leads, in exactly the same way, to $s = v^2$ and $t = 2u^2$. In either case, substituting into $m = s^2 + t^2$ gives

$$v^4 + 4u^4 = s^2 + t^2 = m = a^2.$$

Thus we have another solution of the original equation. Moreover, putting $z_2 = a$, we have

$$0 < z_2 = a \leq m < m^2 + n^2 = z_1,$$

and so the z-value in this new solution is a smaller positive integer.

This completes the descent step and the proof. ■

Corollary 1.10

The Diophantine equation $x^4 - y^4 = z^2$ has no positive solution.

Proof Suppose that x_1, y_1 and z_1 are positive integers with $x_1^4 - y_1^4 = z_1^2$. Squaring both sides of this equation and rearranging gives

$$z_1^4 + 4(x_1 y_1)^4 = (x_1^4 + y_1^4)^2.$$

Thus $x = z_1$, $y = x_1 y_1$, $z = x_1^4 + y_1^4$ is a positive solution of $x^4 + 4y^4 = z^2$. But this contradicts Theorem 1.9 and so the equation $x^4 - y^4 = z^2$ has no positive solution. ■

We finish with a problem for you to attempt. The solution to this problem first appeared in the margin of Fermat's copy of *Diophantus*. We refer to a right-angled triangle associated with a Pythagorean triple as a **Pythagorean triangle**.

Exercise 1.5

Let (x, y, z) be a Pythagorean triple. The area of the associated Pythagorean triangle is $\frac{1}{2}xy$, and for this to be a square we need $xy = 2n^2$ for some integer n.

Show that if there is a positive simultaneous solution to this equation and to $x^2 + y^2 = z^2$, then there is a positive solution of $a^4 - b^4 = c^2$, and hence deduce that no Pythagorean triangle can have an area that is a square.

2 Integral domains

Continuing the discussion at the start of Section 1, we now have proof that the equation $x^4 + y^4 = z^4$ has no integer solutions, and therefore the proof of Fermat's Last Theorem would follow if it could be shown that $x^p + y^p = z^p$ has no solutions where p is an odd prime.

Individual cases were resolved over a number of years: $p = 3$ was proved by Euler in 1770, half a century later $p = 5$ was proved independently by Dirichlet (1805–1859) and Legendre, and $p = 7$ was proved by Lamé (1795–1870) in 1835. Meanwhile, Germain (1776–1831) proved a property about the divisibility of the product xyz of a solution to $x^p + y^p = z^p$, and this enabled her to show that there can be no solution for all odd primes $p \leq 197$.

Lamé thought he had made the breakthrough in 1847. In attempting to solve $x^p + y^p = z^p$, he used the ring

$$\mathbb{Z}[\omega] = \{x_0 + x_1\omega + x_2\omega^2 + \cdots + x_{p-1}\omega^{p-1} : x_i \in \mathbb{Z}\}$$

where ω is a complex primitive pth root of 1, that is, $\omega^p = 1$. Lamé showed that the left-hand side of $x^p + y^p = z^p$ can be factorised as follows:

$$x^p + y^p = (x + y)(x + \omega y)(x + \omega^2 y) \cdots (x + \omega^{p-1} y).$$

Applying Fermat's method of infinite descent, he showed that there can be no trivial solution in this number system, and therefore no positive solution. Unfortunately, this relied on the above factorisation of $x^p + y^p$ being unique, which is not in general true: the equivalent of the Fundamental Theorem of Arithmetic does not always hold in $\mathbb{Z}[\omega]$. For the rest of the chapter, we will be exploring this idea of unique factorisation in general rings, and how this can be used to solve some other Diophantine equations.

In fact, the smallest p for which there is no analogue to the Fundamental Theorem of Arithmetic in $\mathbb{Z}[\omega]$ is $p = 23$.

2.1 Units and associates

Our exploration into unique factorisation begins by picking up the story of rings, zero divisors and units from Chapter 11. For polynomial rings over fields, we have already defined a number of concepts that we will need here, and so some of the terms that we will define should already be familiar. Here, however, we are going to be considering arbitrary commutative rings, and not just polynomial rings. Note that, as in Chapter 11, all the rings we consider will be commutative rings. However, we will occasionally use the term 'commutative' to emphasise the fact that a definition or result does not apply more generally.

We begin by recalling that an element a of a ring R is a **unit** if there exists $b \in R$ for which $a \cdot b = 1$. We saw in Subsection 1.2 of Chapter 11 that the units of $\mathbb{Z}$ are 1 and -1. In other rings, the computation of units can be much harder, as the following example demonstrates.

Example 2.1 *Units*

Consider $\mathbb{Z}[\sqrt{2}]$, the ring of real numbers of the form $a + b\sqrt{2}$ with $a, b \in \mathbb{Z}$. Set $u = 1 + \sqrt{2}$. Then u is a unit, because

$$(1 + \sqrt{2})(-1 + \sqrt{2}) = -1 + \sqrt{2} - \sqrt{2} + \sqrt{2}\sqrt{2} = -1 + 2 = 1.$$

Now, u is a real number that is greater than 1, and so all of the numbers $u, u^2, u^3, \ldots$ are distinct. Moreover, u^k is a unit for every $k \geq 1$, because

$$\begin{aligned}(1 + \sqrt{2})^k(-1 + \sqrt{2})^k &= (1 + \sqrt{2})^{k-1}\left((1 + \sqrt{2})(-1 + \sqrt{2})\right)(-1 + \sqrt{2})^{k-1} \\ &= (1 + \sqrt{2})^{k-1}(-1 + \sqrt{2})^{k-1} \\ &= \cdots = 1.\end{aligned}$$

This shows us, therefore, that $\mathbb{Z}[\sqrt{2}]$ has infinitely many units.

Since $u > 1$, it follows that $1 < u < u^2 < u^3 < \cdots$, which implies that each u^k is distinct. In fact, the only units of $\mathbb{Z}[\sqrt{2}]$ are of the form $1, -1, (1 + \sqrt{2})^k$ or $(-1 + \sqrt{2})^k$, for $k \geq 1$, but we will not prove this here.

How did we end up with infinitely many units in the above example? We managed to find a unit $u = 1 + \sqrt{2}$ for which every power of u is also a unit. This, of course, holds more generally.

Lemma 2.2

Let R be a commutative ring, and u a unit of R. Then u^k is a unit of R for all $k \geq 1$.

Proof Since u is a unit, there exists u^{-1} such that $uu^{-1} = 1$ in R. The inverse of u^k is then $(u^{-1})^k$, since

$$u^k \cdot (u^{-1})^k = u^{k-1}(uu^{-1})(u^{-1})^{k-1} = \cdots = uu^{-1} = 1,$$

and hence u^k is a unit. ■

In $\mathbb{Z}$, we know that the units are ± 1. Now, $(-1)^2 = 1$ (and, of course, $1^2 = 1$), which explains why the above lemma does not mean that $(-1)^k$ creates an infinite list of units in $\mathbb{Z}$.

Knowing about the units of a given ring is important if we are to extend our concepts of factorisation. For example, in $\mathbb{Z}$ we do not want to regard the two factorisations $6 = 2 \cdot 3 = (-2) \cdot (-3)$ as being different: the number -2 is really just the number 2 multiplied by the unit -1, and units are always going to let us do this. This motivates us to extend a concept that we used for polynomial rings (Definition 2.11 in Chapter 11) to a more general setting.

Definition 2.3 *Associate*

Let R be a commutative ring. An **associate** of $a \in R$ is any element that can be written as ua where u is a unit in R.

Exercise 2.1

Let $d > 1$ be a non-square integer.

(a) Show that if $a, b \in \mathbb{Z}$ is a solution to the equation $a^2 - db^2 = 1$, then $a + b\sqrt{d}$ is a unit in $\mathbb{Z}[\sqrt{d}]$.

(b) Hence show that whenever $a^2 - db^2 = 1$ has a solution with $a, b > 0$, then $\mathbb{Z}[\sqrt{d}]$ has infinitely many units.

The equation $a^2 - db^2 = 1$ became known as **Pell's equation** when Euler confused John Pell (1611–1685) with another English mathematician, William Brouncker (1620–1684), who had been studying the equation.

In fact, Pell's equation $a^2 - db^2 = 1$ has infinitely many non-trivial solutions (a, b) for every square-free integer $d > 1$. Although we will not prove this fact here, combining it with the exercise above implies that $\mathbb{Z}[\sqrt{d}]$ has infinitely many units whenever $d > 1$. By contrast, $\mathbb{Z}[\sqrt{-1}]$ has four units, namely $1, -1, i$ and $-i$, and the only units in $\mathbb{Z}[\sqrt{d}]$ for $d < -1$ are -1 and 1.

2.2 Factors, irreducibles, primes and integral domains

When trying to define 'factorisation' for an element $a \in R$, we want to exclude factors that are units, since we can always write $a = uu^{-1}a$ for a unit u and its inverse. Following a similar concept to the one used in our work with polynomial rings, in the next definition we therefore exclude units from being considered proper factors.

We will occasionally use the shorthand notation $a \mid c$ to mean 'a divides c'.

Definition 2.4 *Factor*

Let R be a commutative ring. We say that $a \in R$ is a **factor** or **divisor** of $c \in R$ (or a **divides** c) if there is $b \in R$, such that $c = ab$. If c is non-zero and neither a nor b are units then they are **proper factors** of c.

Note that this definition of factor and proper factor agrees with what we are used to in $\mathbb{Z}$: a proper factor of an integer $n \in \mathbb{Z}$ is any factor of n other than n itself, its associate $-n$ and the two units 1 and -1. Similarly, in the polynomial ring $F[x]$, a proper factor of f must have degree at least one and at most $\deg(f) - 1$, so that it is neither a unit nor an associate.

However, you may have noticed that, unlike in the cases of $\mathbb{Z}$ and $F[x]$, we have defined what it means for one element to divide another without first describing a division algorithm. We will explore why we did this in the next section.

With 'factor' defined, our next step is to generalise the term 'irreducible', which we encountered in Chapter 11. We will present the definition here slightly differently, but it is easy to see that this is equivalent to the approach used in the last chapter.

Definition 2.5 *Irreducible*

Let R be a commutative ring. We say that $a \in R$ is **irreducible** if it is not a unit, it is not zero and whenever $a = bc$ for $b, c \in R$, then one of b or c is a unit.

In the integers, we used the term 'prime' instead of irreducible, and you may be wondering why the terminology has changed. The reason is that in the general context of studying rings, the term 'prime' is used to define something related, but subtly different.

Definition 2.6 *Prime*

Let R be a commutative ring. A non-zero, non-unit element p of R is **prime** if whenever $p \mid ab$ for $a, b \in R$, then either $p \mid a$ or $p \mid b$.

We will soon see that Definitions 2.5 and 2.6 coincide in $\mathbb{Z}$ (and define the 'prime numbers' that we are familiar with). In the more general setting of rings, understanding the distinction between irreducible and prime is key to our exploration of unique factorisation. When do they define different objects? There are really two questions here: when are irreducible elements prime, and when are prime elements irreducible? We will consider the second of these questions in the next subsection. To answer the first, we consider an example.

Example 2.7 *A prime element that is not irreducible*

Take the ring $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$, of integers with addition and multiplication modulo 6. Then the element 2 is prime in $\mathbb{Z}_6$. To see this, we write out the multiplication table for $\mathbb{Z}_6 \setminus \{0\}$.

	1	2	3	4	5
1	1	2	3	4	5
2	2	4	0	2	4
3	3	0	3	0	3
4	4	2	0	4	2
5	5	4	3	2	1

Reading along the row for 2, we see that 2 divides the numbers 2, 4 and 0. Now, we can write 4 as the product of two elements in four ways:

$$4 = 1 \cdot 4 = 2 \cdot 2 = 2 \cdot 5 = 4 \cdot 4.$$

In each of these four products, we find that 2 divides one of the two factors. We can do the same with 2:

$$2 = 1 \cdot 2 = 2 \cdot 4 = 4 \cdot 5$$

and with 0:

$$0 = 2 \cdot 3 = 3 \cdot 4 = 0 \cdot c = c \cdot 0, \quad \text{for } c \in \mathbb{Z}_6.$$

Thus, whenever 2 divides a product of two elements ab, then 2 divides a or b, and so is prime.

However, 2 is not irreducible: for example $2 = 2 \cdot 4$, but neither 2 nor 4 are units.

We also know $2 = 4 \cdot 5$. However, $5 \cdot 5 = 1$ and so 5 is a unit. This, therefore, does not help us to prove that 2 is not irreducible.

Exercise 2.2

(a) List the units, irreducible and prime elements in (i) $\mathbb{Z}_6$ and (ii) $\mathbb{Z}_8$.

(b) Is $4 + 3i$ irreducible in $\mathbb{Z}[i]$, the ring of Gaussian integers? (Do not spend too long on this before referring to the solution.)

In the solution to part (b) of the above exercise, we suggested that the **complex modulus** $|a + bi| = \sqrt{a^2 + b^2}$ might be used to find a factorisation of the complex number $a + bi$. This approach will be invaluable later in the chapter.

In Example 2.7, what was it about $\mathbb{Z}_6$ that made the number 2 reducible? First, notice that 2 is a zero divisor in $\mathbb{Z}_6$, since $2 \cdot 3 = 0$. Using the basic axioms for a ring, we can now write 2 as a product of two proper factors:

You may wish to remind yourself how zero divisor was defined in Definition 1.10 of Chapter 11.

$$\begin{aligned} 2 &= 0 + 2 \\ &\equiv (2 \cdot 3) + (2 \cdot 1) \\ &\equiv 2 \cdot (3 + 1) \equiv 2 \cdot 4 \,(\text{mod } 6). \end{aligned}$$

However, here we should add a word of caution: in Exercise 2.2(a)(ii), you should have found that in $\mathbb{Z}_8$ the number 2 is both prime and irreducible, even though it is also a zero divisor. Following the same argument as for $\mathbb{Z}_6$, we have

$$2 = 0 + 2 \equiv (2 \cdot 4) + (2 \cdot 1) \equiv 2 \cdot 5 \,(\text{mod } 8),$$

and we get the factorisation $2 = 2 \cdot 5$. However, 5 is a unit in $\mathbb{Z}_8$, and so it is not a proper factor.

Thus, in some rings, simply being a zero divisor might not be enough to give us a prime non-irreducible element. However, we can escape these complexities by restricting our attention to rings with no zero divisors.

Definition 2.8 *Integral domain*

A commutative ring is an **integral domain** if it has no zero divisors.

Equivalently, a ring R is an integral domain if for $a, b \in R$, whenever $a \cdot b = 0$ then either $a = 0$ or $b = 0$.

Example 2.9 *Some integral domains*

(a) As we should expect, $\mathbb{Z}$ is an integral domain.

(b) Any field F is an integral domain by Lemma 1.17 of Chapter 11. Therefore, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all integral domains.

(c) Any subring of a field is also an integral domain.

In fact, every integral domain is a subring of some field. If R is an integral domain, the smallest field F for which R is a subring of F is called the **field of fractions**. We will encounter these again in Book E.

Example 2.10 *Some rings that are not integral domains*

(a) When $n \in \mathbb{N}$ is composite, the ring $\mathbb{Z}_n$ contains zero divisors, as was shown in Exercise 1.8 of Chapter 11.

(b) The ring $\mathbb{Z} \times \mathbb{Z} = \{(a, b) : a, b \in \mathbb{Z}\}$ also contains zero divisors. This ring was considered in Exercise 1.6 of Chapter 11.

The key property of integral domains is the ability to 'cancel' terms, which gives us the following equivalent way to define an integral domain.

Lemma 2.11 *Cancellation in an integral domain*

Let R be a commutative ring. Then R is an integral domain if, and only if, for $a, b, c \in R$ with c non-zero, $ac = bc$ implies $a = b$.

Proof First, assume R is an integral domain. We will show that if $ac = bc$ for a non-zero element c then $a = b$. From $ac = bc$ we get $0 = ac - bc = (a - b)c$. Now, since R is an integral domain and $c \neq 0$, we must have $a - b = 0$ and therefore $a = b$.

Note that $-bc = (-b)c$ by Proposition 1.5 of Chapter 11.

Conversely, suppose R is not an integral domain. Then there exist non-zero $a, b \in R$ such that $ab = 0$. We now have $a \cdot 0 = 0 = ab$. However, $b \neq 0$, and so $a0 = ab$ does not imply $0 = b$. ■

Reassuringly, we now find that every prime element is irreducible in an integral domain.

Theorem 2.12

Let R be an integral domain and let p be a prime element of R. Then p is irreducible.

Proof Suppose $p = ab$ for some $a, b \in R$. We need to show that one of a or b is a unit. Since p is prime, and $p \mid ab$, either $p \mid a$ or $p \mid b$. Without loss of generality, assume that $p \mid a$. Then $a = pq$ for some $q \in R$ and so $p = ab = pqb$. Now $p = pqb$, and by Lemma 2.11, this implies that $1 = qb$ and so b is a unit. ■

Now that we are certain that restricting our attention to integral domains is a step in the right direction, we can list some useful properties of division, all of which follow directly from the definition.

Lemma 2.13 *Properties of division in an integral domain*

Let R be an integral domain and let a, b, c be non-zero elements in R.

(a) If $a \mid c$, then $a \mid (c + na)$ for any $n \in R$.

(b) If $a \mid b$ and $b \mid a$, then a is an associate of b.

(c) If $a \mid b$ and $b \mid c$, then $a \mid c$.

(d) If $a \mid b$ and $a \mid c$, then $a \mid (mb + nc)$ for any $m, n \in R$.

(e) Every element of R divides 0.

(f) If u is a unit in R, then $u \mid n$ for every $n \in R$.

Exercise 2.3

Prove Lemma 2.13.

Exercise 2.4

Prove that if S is a subring of an integral domain R, then S is an integral domain.

2.3 Norms

In the last subsection we established that prime elements need not be irreducible in an arbitrary ring, but this is not the case in integral domains. In this subsection, we will begin to consider the question of whether irreducible elements are always prime.

First, let us consider the ring of integers $\mathbb{Z}$. Translating into the language of this chapter, Euclid's Lemma for prime factors (Theorem 1.4 of Book A, Chapter 2) tells us that an irreducible element in $\mathbb{Z}$ satisfies Definition 2.6. Coupling the above with Theorem 2.12, we find that in $\mathbb{Z}$, the definitions of 'prime' and 'irreducible' do in fact coincide.

Will prime and irreducible always be the same in an integral domain, or do we need more conditions to ensure this? To investigate this question, we first need some theory.

Definition 2.14 *Norm*

Let R be an integral domain. A **norm** is a function

$$N : R \setminus \{0\} \to \mathbb{N} \cup \{0\},$$

mapping the non-zero elements of R to the non-negative integers.

As we will see later, we need Euclidean norms to generalise the Division Algorithm.

N is a **Euclidean norm** if for non-zero $a, b \in R$ we have

(a) $N(a) \leq N(ab)$, and

(b) there exist $q, r \in R$ such that $a = bq + r$, and either $r = 0$ or $N(r) < N(b)$.

N is a **multiplicative norm** if for non-zero $a, b \in R$ we have $N(ab) = N(a)N(b)$.

The idea behind these definitions is that they will give us a 'measure' of the relative sizes of elements of R. Let us look at some examples.

Example 2.15 *Some norms*

(a) In the ring of integers $\mathbb{Z}$, the usual norm of an element $a \in \mathbb{Z} \setminus \{0\}$ is $|a|$, the absolute value of a.

It is multiplicative since $|ab| = |a| \cdot |b|$, and the Division Algorithm shows that it is a Euclidean norm.

(b) For a polynomial ring $F[x]$ over a field F, the degree of a non-zero polynomial is a norm, $N(f) = \deg(f)$.

It is not multiplicative, since $\deg(fg) = \deg(f) + \deg(g)$, but we do have $\deg(f) \leq \deg(fg)$ and so the Division Algorithm for polynomials (Theorem 3.1 of Chapter 11) shows that it is Euclidean.

(c) In the Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$, the complex modulus squared, $N(a + bi) = a^2 + b^2$, is a norm. N is multiplicative since $(a + bi)(c + di) = (ac - bd) + (ad + bc)i$, and so we have

$$\begin{aligned} N((a + bi)(c + di)) &= (ac - bd)^2 + (ad + bc)^2 \\ &= (a^2 + b^2)(c^2 + d^2) \\ &= N(a + bi)N(c + di). \end{aligned}$$

We discuss this norm further in Worked Exercise 3.4.

We can now return to the question of whether irreducible elements need always be prime in an integral domain. In our investigation, we will actually also encounter our first example of non-unique factorisation.

Consider the ring $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$. In this ring, we can write the element 6 in two different ways:

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

This factorisation in two different ways *looks* like it should give us an example: assuming the element $2 \in \mathbb{Z}[\sqrt{-5}]$ is irreducible, then $2 \mid 6$, but it looks unlikely that 2 is going to divide either $(1 + \sqrt{-5})$ or $(1 - \sqrt{-5})$. If this really is true (and remember, we haven't proved anything yet), then this will show us that 2 is not prime.

Moreover, if the elements $3, (1 + \sqrt{-5})$ and $(1 - \sqrt{-5})$ are also all irreducible and none is an associate of another, then this will give us an example of non-unique factorisation.

What does it mean for an element to have 'unique factorisation'? We will define this more formally later, but it should mean that it can be uniquely expressed as a product of irreducible elements.

We have just made rather a lot of assumptions about elements in $\mathbb{Z}[\sqrt{-5}]$, and so now we need to prove these. The following exercise will lead you through the formal details – we will use a norm function to help us to prove that elements are irreducible.

Exercise 2.5 *Finding an irreducible element that is not prime*

Let R be the ring $\mathbb{Z}[\sqrt{-5}]$.

(a) Show that R is an integral domain.

(b) Define a norm $N : R \setminus \{0\} \to \mathbb{N} \cup \{0\}$ by $N(a + b\sqrt{-5}) = a^2 + 5b^2$. Show that N is a multiplicative norm.

(c) Show that if $r = a + b\sqrt{-5} \neq 0$ and $b \neq 0$ then $N(r) \geq 5$. What can you say about r if $N(r) < 5$?

(d) Why is 2 an irreducible element of R?
Hint: what is $N(2)$?

(e) Show that 2 does not divide $1 + \sqrt{-5}$ or $1 - \sqrt{-5}$, and hence show that 2 is not prime in R.

Exercise 2.6 *Proving non-unique factorisation*

As in Exercise 2.5, let R be the ring $\mathbb{Z}[\sqrt{-5}]$, and let N be the multiplicative norm $N(a+b\sqrt{-5}) = a^2+5b^2$.

(a) Show that if u is a unit of R then $N(u) = 1$, and deduce that the only units of R are $1, -1$.

(b) Show that 3 is irreducible in R.

(c) Show that $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are irreducible elements of R.

(d) Hence factorise 6 as a product of irreducible elements in two different ways in R.

3 Euclidean domains

To explore uniqueness of factorisation further, this section uses the Euclidean norm from the previous section to identify a family of integral domains that have a division algorithm.

3.1 Rings with a division algorithm

We saw in the last section that there are rings (and, indeed, integral domains) where irreducible elements need not be prime, and this allowed us to construct an example where factorisation into irreducibles is not unique. In order to isolate the reasons why this can happen, let us compare what we know for general rings with what we know for $\mathbb{Z}$, where uniqueness of factorisation does hold.

Look back for a moment at Euclid's Lemma for prime factors, Theorem 1.4 in Book A, Chapter 2. Translating into our language, this says 'if $p \in \mathbb{Z}$ is irreducible, and p divides the product $a_1a_2\cdots a_n$, then p divides a_i for some i'. In other words, in $\mathbb{Z}$, if p is irreducible then it is prime – this is exactly what we want to see. How did we prove this, and can we mimic this proof in a more general setting?

Following the proof of Theorem 1.4, we find that we need to know about 'highest common factors' and be certain that these exist. In $\mathbb{Z}$ this was Proposition 4.7 of Book A, Chapter 1, and central to this proof was the Division Algorithm. This is what we need to generalise, and we have some experience of this already, through our study of polynomial division in Chapter 11.

For $\mathbb{Z}$, the Division Algorithm (Theorem 4.1 of Book A, Chapter 1) means that any integer a can be divided by any positive integer b to leave a 'small' remainder, that is, $a = bq + r$ where $r, q \in \mathbb{Z}$, and r is in the range $0 \leq r < b$. For polynomials with coefficients in a field, we defined a division algorithm where 'smaller' meant of lower degree (Theorem 3.1 of Chapter 11). To measure 'smallness' in arbitrary rings, we can use the ready-made concept of a norm, which leads us to the following definition.

Definition 3.1 *Euclidean domain*

Let R be an integral domain. If there exists a Euclidean norm $N : R \setminus \{0\} \to \mathbb{N} \cup \{0\}$ on R, then we say that R is a **Euclidean domain**.

Recall that a Euclidean norm satisfies two conditions. For non-zero $a, b \in R$:

(i) $N(a) \leq N(ab)$

(ii) there exist $q, r \in R$ such that $a = bq + r$ where $N(r) < N(b)$ or $r = 0$.

Example 3.2 *Some Euclidean domains*

(a) The ring of integers $\mathbb{Z}$ is a Euclidean domain with norm $N(a) = |a|$ for $a \in \mathbb{Z} \setminus \{0\}$. We observed in Example 2.15(a) that this norm is Euclidean, which follows from the results of Book A, Chapter 1.

(b) Any field F is a Euclidean domain. Since $F \setminus \{0\}$ consists entirely of units, for any $a, b \in F \setminus \{0\}$ we can take $q = b^{-1}a \in F$ so that $a = bq$. In other words, we can always take remainder 0, and therefore our norm function needs to satisfy only $N(a) \leq N(ab)$. Trivially, we may set $N(a) = 1$ for all $a \in F \setminus \{0\}$.

(c) The polynomial ring $F[x]$ where F is a field and $N(f) = \deg(f)$ for non-zero $f \in F[x]$ is a Euclidean domain. This was proved in Chapter 11.

Notice that it is possible for a ring to be a Euclidean domain with respect to more than one norm. For example, the ring of integers is also a Euclidean domain with respect to the norm $N(a) = a^2$.

From the definition of Euclidean norm, we immediately derive the following division algorithm.

Theorem 3.3 *Division Algorithm for Euclidean domains*

Let R be a Euclidean domain with Euclidean norm N and let $a, b \in R$ with b non-zero. Then there exist $q, r \in R$ such that $a = bq + r$ with $N(r) < N(b)$ or $r = 0$.

The following worked exercise gives an example of how to establish that a given function is a Euclidean norm and shows how to find elements q and r as required for the Division Algorithm.

Worked Exercise 3.4

Prove that the ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain with $N(x+yi) = x^2 + y^2 = |x+yi|^2$.

Solution

First, $\mathbb{Z}[i]$ is an integral domain since it is a subring of $\mathbb{C}$, and so it remains to show that N is a Euclidean norm.

For non-zero $a, b \in \mathbb{Z}[i]$,

$$N(ab) = N(a)N(b) \geq N(a)$$

since $N(b) \geq 1$ for all non-zero $b \in \mathbb{Z}[i]$.

Finally, for $a, b \in \mathbb{Z}[i]$ and $b \neq 0$, we need to find $q, r \in \mathbb{Z}[i]$ such that $a = bq + r$ and $N(r) < N(b)$. We let q be a point in $\mathbb{Z}[i]$ that is nearest to the complex number $z = \frac{a}{b}$ (Figure 3.1). The point z lies inside a 1×1 square, and so the distance from z to the nearest corner of the square (that is, the point q) is at most $\frac{\sqrt{2}}{2}$. In other words, the distance from z to q, $|z-q|$, satisfies

$$|z-q| \leq \frac{\sqrt{2}}{2} = \frac{1}{\sqrt{2}} < 1.$$

Now, let $r = a - bq$. Then $r \in \mathbb{Z}[i]$, and

$$N(r) = |r|^2 = |a-bq|^2 = |b|^2|z-q|^2 < |b|^2 = N(b),$$

which proves that N is a Euclidean norm.

Note that q might not be unique, but we are really interested in the distance from z; so any closest point q will do.

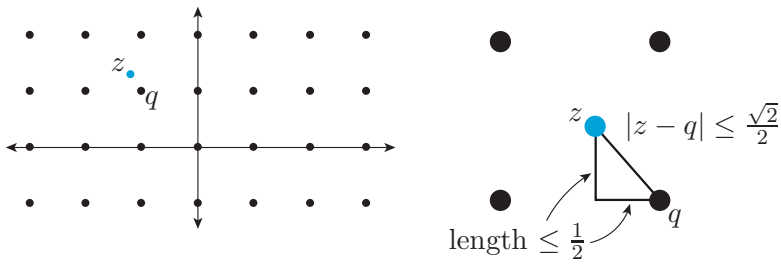


Figure 3.1 A nearest point $q \in \mathbb{Z}[i]$ to the point $z \in \mathbb{C}$

Exercise 3.1

(a) Following the approach in Worked Exercise 3.4, prove that $\mathbb{Z}[\sqrt{-2}]$ is a Euclidean domain, with norm $N(a+b\sqrt{-2}) = a^2+2b^2 = |a+b\sqrt{2}i|^2$.

(b) Explain why a similar argument cannot be used for $\mathbb{Z}[\sqrt{-5}]$ with the norm $N(a+b\sqrt{-5}) = a^2+5b^2$.

(c) By considering the elements 2 and $1+\sqrt{-5}$, show that $\mathbb{Z}[\sqrt{-5}]$ is *not* a Euclidean domain with respect to the norm $N(a+b\sqrt{-5}) = a^2+5b^2$.

Before we use our newly formed Division Algorithm to advance our story, let us pause briefly to record some basic properties of Euclidean norms.

Lemma 3.5

Let R be a Euclidean domain with norm N. Then:

(a) $N(1)$ is the smallest value taken by the function N over all non-zero elements of R.

(b) $N(u) = N(1)$ if, and only if, u is a unit in R.

(c) If a, b are associates in R then $N(a) = N(b)$.

(d) For any non-zero $a \in R$, if $b \neq 0$ is not a unit then $N(a) < N(ab)$.

Proof

(a) For any non-zero element $b \in R$, we have $N(b \cdot 1) \geq N(1)$ from Definition 2.14 and so $N(b) \geq N(1)$.

(b) Suppose that u is a unit in R. Then $N(u) \leq N(uu^{-1}) = N(1)$ $\leq N(u \times 1) = N(u)$, and so $N(u) = N(1)$.

Conversely, suppose $u \neq 0 \in R$ and $N(u) = N(1)$. Then there exist $q, r \in R$ such that $1 = uq + r$ with either $N(r) < N(u) = N(1)$ or $r = 0$. Since we cannot have $N(r) < N(1)$ by part (a), it follows that $r = 0$ and $1 = uq$. In other words, u is a unit in R.

(c) Since a and b are associates, there is a unit $u \in R$ so that $au = b$. Then $N(b) = N(au) \geq N(a)$ and $N(a) = N(bu^{-1}) \geq N(b)$. Hence $N(a) = N(b)$.

(d) Suppose that $a, b \in R$ are non-zero and that b is not a unit. Since N is Euclidean, we have $N(a) \leq N(ab)$. If $N(a) = N(ab)$, we want to conclude that b is a unit, which will be a contradiction. Since a and ab are non-zero, we can find q, r such that

$$a = (ab)q + r,$$

where either $r = 0$ or $N(r) < N(ab) = N(a)$. Rearranging, we have $r = a(1-bq)$. We cannot have $r \neq 0$, as this would imply $N(r) = N(a(1-bq)) \geq N(a)$, which is impossible. Therefore $a(1-bq) = 0$, and since $a \neq 0$ we must have $1 - bq = 0$. This means that $bq = 1$, and so b is a unit, as required. ■

The above lemma gives us a method of computing units, as the next exercise shows.

Exercise 3.2

(a) Prove that the only units in $\mathbb{Z}[i]$ are $\pm 1, \pm i$.

(b) Compute the units in the Euclidean domain $\mathbb{Z}[\sqrt{-2}]$, with norm $N(a + b\sqrt{-2}) = a^2 + 2b^2$.

3.2 Highest common factors

We begin this subsection with a direct generalisation of the definition of highest common factor, to apply to an integral domain.

Definition 3.6 *Highest common factor*

Let a and b be elements of an integral domain R. An element h is said to be a **highest common factor** (**hcf**) of a and b if it satisfies the following:

(a) $h \mid a$ and $h \mid b$

(b) if $d \mid a$ and $d \mid b$, then $d \mid h$.

Highest common factors are also known as **greatest common divisors**, denoted by **gcd**.

Condition (a) says that a highest common factor *is* a common factor of elements, while condition (b) says that it is *maximal* amongst the common factors. Note that we can talk only about *a* (rather than *the*) highest common factor because if $h = \text{hcf}(a, b)$, then any associate $u \cdot h$ (where u is a unit) also satisfies conditions (a) and (b). In our definition of highest common factor for the integers we specified that we would take the positive number to ensure uniqueness, and for polynomial rings $F[x]$ over fields we fixed the highest common factor to be the unique monic polynomial, but in this more general setting we cannot meaningfully enforce such a rule, and so we have to accept non-uniqueness.

As well as non-uniqueness, there also remains the question of whether hcfs must exist. First, as we did in $\mathbb{Z}$ and $F[x]$, we adopt the convention that $\text{hcf}(0, 0)$ never exists, and so we should consider $\text{hcf}(a, b)$ only where at least one of a or b is non-zero. In fact, considering only integral domains, there are still instances where hcfs do not exist.

Worked Exercise 3.7

Consider the ring $\mathbb{Z}[\sqrt{-5}]$. Prove that the elements 6 and $2(1+\sqrt{-5})$ do not have a highest common factor.

Solution

We will make use of the multiplicative norm $N(a+b\sqrt{-5}) = a^2+5b^2$ (introduced in Exercise 2.5). First, $N(6) = 36$ and $N(2(1+\sqrt{-5})) = 4+20 = 24$, and so if $h = \text{hcf}(6, 2(1+\sqrt{-5}))$ exists then $N(h)$ must divide 24 and 36. Thus $N(h)$ must be equal to one of 1, 2, 3, 4, 6 or 12.

As we saw in Exercise 2.5, the element 6 can be written as the product of irreducibles in two different ways:

$$6 = 2\cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5}).$$

Consequently, 6 and $2(1+\sqrt{-5})$ have both 2 and $1+\sqrt{-5}$ as common factors, and so both of these must divide h. Now $N(2) = 4$ and $N(1+\sqrt{-5}) = 6$, and so 4 and 6 must both divide $N(h)$, which implies that the only possibility is that $N(h) = 12$.

Writing $h = a+b\sqrt{-5}$, we have $N(h) = a^2+5b^2 = 12$. From this, the only options are $b = 0$ or 1, for otherwise $a^2+5b^2 \geq 20 > 12$. When $b = 0$ we have $a^2 = 12$, while if $b = 1$ then $a^2 = 12-5 = 7$; neither of these has solutions in $\mathbb{Z}$. Therefore h cannot exist.

When we proved the existence of hcfs in the integers (and in polynomial rings over fields), we repeatedly applied the Division Algorithm until the remainder was 0. We were guaranteed that this repeated application of the Division Algorithm would stop, because in the equation $a = bq+r$ we ensured $0 \leq r < b$. For Euclidean domains, the equivalent condition is 'either $N(r) < N(b)$ or $r = 0$', and this gives us the tool we need to define the **Euclidean Algorithm** (and consequently to prove the existence of hcfs) for arbitrary Euclidean domains.

Theorem 3.8

Let R be a Euclidean domain with norm N, and let $a, b \in R$ with b non-zero. Then a and b have a highest common factor. If $h = \text{hcf}(a,b)$, then there are elements $s, t \in R$ so that $h = sa+tb$.

Proof We begin by defining $S = \{xa + yb : x, y \in R\}$, and notice that S is non-empty because $a \in S$ and $b \in S$, as shown by $a = 1 \cdot a + 0 \cdot b$ and $b = 0 \cdot a + 1 \cdot b$.

Now, by the Well-Ordering Principle from Book A, Chapter 1, we choose non-zero $h \in S$ such that $N(h)$ is as small as possible. Since $h \in S$, we know that $h = sa + tb$ for some $s, t \in R$.

It remains to show that h is indeed a highest common factor of a and b. First, if $d \mid a$ and $d \mid b$, then clearly $d \mid xa + yb$ for any $x, y \in R$, and so $d \mid h$. Since R is a Euclidean domain, by Definition 3.1 we have $a = qh + r$ for some $q, r \in R$, with either $r = 0$ or $N(r) < N(h)$. Now, if $r \neq 0$ then

$$r = a - qh = a - q(sa + tb) = (1 - qs)a - (qt)b \in S.$$

We now have a contradiction because $r \in S$ and $N(r) < N(h)$, but h was chosen to minimise $N(h)$ over all non-zero elements in S. Therefore $r = 0$, $a = qh$, and consequently $h \mid a$. A similar argument can be applied to show that $h \mid b$. ■

In the solution to Exercise 3.1(c), we showed that $\mathbb{Z}[\sqrt{-5}]$ was not a Euclidean domain with respect to the norm $N(a + b\sqrt{-5}) = a^2 + 5b^2$. Could it be a Euclidean domain with respect to some other norm? Theorem 3.8 now gives us a way of showing that the answer to this question is no.

Worked Exercise 3.9

Show that $\mathbb{Z}[\sqrt{-5}]$ is not a Euclidean domain with respect to any norm.

Solution

We proved (in Worked Exercise 3.7) that the elements 6 and $2(1 + \sqrt{-5})$ have no highest common factor in $\mathbb{Z}[\sqrt{-5}]$. Therefore $\mathbb{Z}[\sqrt{-5}]$ cannot be a Euclidean domain with respect to any norm, by Theorem 3.8.

Note that even though we specified a norm to prove that 6 and $2(1 + \sqrt{-5})$ have no hcf, the hcf itself is defined without reference to any norm.

Theorem 3.8 tells us not only that highest common factors exist in Euclidean domains, but that $\text{hcf}(a, b)$ must be expressible as a linear combination of a and b. This additional requirement gives us another method to prove that some rings are not Euclidean domains, as the next exercise shows.

Exercise 3.3

(a) Show that $\mathrm{hcf}(2, x) = 1$ in $\mathbb{Z}[x]$.

(b) Hence, or otherwise, conclude that $\mathbb{Z}[x]$ is not a Euclidean domain.

(c) Find polynomials $s, t \in \mathbb{Q}[x]$ such that $2s(x) + xt(x) = 1$ in $\mathbb{Q}[x]$.

The existence of highest common factors in Euclidean domains now gives us a route to what we have been working towards in this section: situations where irreducible and prime mean the same thing. First, we need a preliminary result.

Lemma 3.10

Suppose that p is an irreducible element of a Euclidean domain R, that a is a non-zero element of R and that p does not divide a. Then there exist $s, t \in R$ with $ps + at = 1$.

Proof Let b be a highest common factor of a and p. Then by Theorem 3.8 there exist $s', t' \in R$ with $b = s'a + t'p$. Now since b divides p and p is irreducible, either $b = up$ where u is a unit, or b is a unit. If $b = up$ then, since b divides a, p also divides a. Thus b is a unit and so $1 = b^{-1}b = b^{-1}s'a + b^{-1}t'p$ and the result follows with $s = b^{-1}s'$ and $t = b^{-1}t'$. ■

We are now ready for the following result.

Proposition 3.11

Let R be a Euclidean domain with norm N, and let p be an irreducible element in R. Then p is prime in R.

Proof Let $a, b \in R$ be such that p divides ab. If p divides a, then we have finished; so let us suppose that p does not divide a. By Lemma 3.10, there exist $s, t \in R$ such that

$$ps + at = 1.$$

Multiplying this equation by b, we obtain

$$pbs + abt = b.$$

Now, since $p \mid ab$, and p must divide itself, we can apply Lemma 2.13(d) to show that p divides $pbs + abt$. Since $b = pbs + abt$, we finally conclude that p must divide b. Hence p is prime. ■

Recall that to prove p is prime, we need to show that whenever $p \mid ab$, then $p \mid a$ or $p \mid b$.

3.3 Prime numbers expressed as the sum of two squares

Before we continue our exploration of unique factorisation, let us pause for a moment to present a number theoretic result that follows from our study of Euclidean domains.

Another problem that attracted Fermat's attention concerned ways of expressing positive integers as sums of squares. For example, the integers from 1 to 9 can be written as sums of squares as follows.

$$1 = 1^2$$
$$2 = 1^2 + 1^2$$
$$3 = 1^2 + 1^2 + 1^2$$
$$4 = 2^2$$
$$5 = 2^2 + 1^2$$
$$6 = 2^2 + 1^2 + 1^2$$
$$7 = 2^2 + 1^2 + 1^2 + 1^2$$
$$8 = 2^2 + 2^2$$
$$9 = 3^2$$

The expressions given are certainly not unique; for example, we could write

$$4 = 1^2 + 1^2 + 1^2 + 1^2 \quad \text{or} \quad 9 = 2^2 + 2^2 + 1^2.$$

But notice that the given expression for 7 as a sum of squares is the best we can do in the sense that 7 cannot be expressed as a sum of fewer than four squares. If you try continuing the list that we have started, with the aim of expressing each integer as a sum of as few squares as possible, you should find that each requires no more than four squares.

Amongst other questions that these observations might suggest are the following.

We accept 0 as a square so that, for example, $4 = 2^2 + 0^2$ is a legitimate way of expressing 4 as a sum of two squares. However, we ignore negative integers since $(-a)^2 = a^2$.

- Which integers can be expressed as a sum of two squares?
- Which integers can be expressed as a sum of two squares in a unique way?
- Which integers can be expressed as a sum of three squares?
- Can all positive integers be expressed as a sum of four squares?

In this subsection, we begin an investigation into the two-squares problem. The following identity, which is easily established, is going to be crucial in what follows.

Important identity for two squares

$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2$$

This identity should look familiar: we used it when proving that the complex modulus, $|a + bi|^2 = a^2 + b^2$, is a multiplicative norm in Example 2.15(c) of this chapter.

This identity is of theoretical importance because it tells us that if two positive integers m and n can each be written as a sum of two squares then

so too can their product mn. So in order to show that a given integer can be written as a sum of two squares, it is sufficient to show that each prime in its decomposition can be expressed in this way.

But the identity also has a practical use, as illustrated by the following example.

Worked Exercise 3.12

Express 4420 as a sum of two squares.

Solution

$$4420 = 2^2 \times 5 \times 13 \times 17 = (2^2+0^2)(2^2+1^2)(3^2+2^2)(4^2+1^2)$$

Using the identity three times:

$$\begin{aligned} 4420 &= (4^2+2^2)(3^2+2^2)(4^2+1^2) \\ &= (16^2+2^2)(4^2+1^2) \\ &= 66^2+8^2. \end{aligned}$$

This answer is not unique. It turns out that there are four ways of writing 4420 as a sum of two squares. The other ways can all be obtained from the identity applied to different terms. We could change the order of the bracketed terms, or switch the two terms within a pair of brackets, or make use of negative values for a, b, c or d. Each of these variations is illustrated below, where we discover the other three solutions.

$$\begin{aligned} 4420 &= (4^2+2^2)(4^2+1^2)(2^2+3^2) \\ &= (4^2+2^2)(11^2+10^2) = 64^2+18^2 \\ 4420 &= (4^2+2^2)(11^2+(-10)^2) = 24^2+(-62)^2 = 62^2+24^2 \\ 4420 &= ((4^2+1^2)(2^2+1^2))((2^2+0^2)(2^2+3^2)) \\ &= (9^2+2^2)(4^2+6^2) = 48^2+46^2 \end{aligned}$$

Exercise 3.4

Find three ways of expressing 325 as a sum of two squares.

Turning to the question of whether a positive integer n can be expressed as a sum of two squares, our important identity guides us to look at the primes occurring in its decomposition. Now primes can be classified into three types:

- the even prime $2 = 1^2 + 1^2$, which is a sum of two squares
- odd primes of the form $4k + 1$
- odd primes of the form $4k + 3$.

The third category presents little problem. As a square is congruent modulo 4 to either 0 or 1, a sum of two squares is congruent modulo 4 to one of 0, 1 or 2. So no prime of the form $4k + 3$ can be expressed as a sum of two squares.

That leaves just the second category. We will prove that all primes of the form $4k + 1$ can be written as a sum of two squares. The original proof, due to Fermat, used the infinite descent method that we encountered in Section 1, albeit in a slightly different way. The (shorter) proof that we present here makes use of the fact that the ring of Gaussian integers, $\mathbb{Z}[i]$, is a Euclidean domain with respect to the norm $N(a + bi) = |a + bi|^2 = a^2 + b^2$.

Theorem 3.13

A prime number p can be expressed as the sum of two squares if, and only if, it is either even, or odd of the form $4k + 1$ for some integer k.

Proof First, if $p = u^2 + v^2$ is an odd number, then (by the comments made before the statement of the theorem) $p = 4k + 1$, which, together with noting that $2 = 1 + 1$ is the sum of two squares, proves the 'only if' part.

For the other direction, we first of all recall Theorem 2.2 from Chapter 10: since $p \equiv 1 \pmod 4$, we have the Legendre symbol $(-1/p) = 1$, that is, -1 is a square modulo p. This means that there must exist $a \in \mathbb{Z}_p$ such that $a^2 \equiv -1 \pmod p$. Thus, p divides $a^2 + 1$.

We now start working in $\mathbb{Z}[i]$: first, note that $a^2 + 1 = (a + i)(a - i)$, and so if p is prime in $\mathbb{Z}[i]$ then p must divide either $a + i$ or $a - i$. This, however, is impossible as p does not divide the imaginary part of either $a + i$ or $a - i$. Therefore, p is not prime and hence, since $\mathbb{Z}[i]$ is a Euclidean domain, p is not irreducible in $\mathbb{Z}[i]$ by Proposition 3.11. Therefore we can write $p = (a + ib)(c + id)$ as a product of two non-units in $\mathbb{Z}[i]$.

If $p \mid a + i$, then
$$a + i = p \cdot (b + ci) = pb + (pc)i$$
for some b, c in $\mathbb{Z}$. But then $pc = 1$, which in $\mathbb{Z}$ is impossible since $p > 1$.

Taking norms, we have $p^2 = (a^2 + b^2)(c^2 + d^2)$, and as $a + ib$ and $c + id$ are not units, both $a^2 + b^2$ and $c^2 + d^2$ must be greater than 1. It follows from the Fundamental Theorem of Arithmetic that $p = a^2 + b^2$ is the sum of two squares. ■

Knowing that a prime $p = 4k + 1$ can be expressed as $x^2 + y^2$, mathematicians understandably took up the challenge of finding ways of constructing the integers x and y in terms of the prime p. Several such constructions have emerged, but none of them is particularly easy to employ.

For practical purposes, the method of exhaustion offers as good a method as any for expressing moderately sized primes as a sum of two squares. For instance, consider the prime 277. First note that if $277 = x^2 + y^2$ then one of x^2 or y^2 is less than $\frac{277}{2}$, while the other is greater than $\frac{277}{2}$. By virtue of Theorem 3.13 we need to check

$$277 - 1^2, \quad 277 - 2^2, \quad 277 - 3^2, \quad \ldots$$

only until we find a square, in the certain knowledge that one will turn up by the time we reach $277 - 12^2$, since $12^2 > \frac{277}{2}$. In fact $277 - 9^2 = 14^2$.

To represent a composite number as a sum of two squares, we can combine the suggested exhaustive approach with use of the important identity.

Exercise 3.5

Express $5321 = 17 \times 313$ as a sum of two squares.

4 Unique factorisation domains

This section takes a direct approach to unique factorisation by naming integral domains that have this property 'unique factorisation domains'.

4.1 Prime and irreducible in unique factorisation domains

In Exercise 2.5 of this chapter, we found that there are irreducible elements in $\mathbb{Z}[\sqrt{-5}]$ that are not prime. In Exercise 2.6, we used the irreducible non-prime elements from Exercise 2.5 to establish that the element 6 of $\mathbb{Z}[\sqrt{-5}]$ can be written as a product of irreducibles in more than one way:

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

We now explore rings where this is not the case for any element. We begin by clarifying exactly what we mean when we say that a ring has unique factorisation. The complications in this definition arise solely to handle units. For example, writing $6 = 2 \cdot 3 = (-2) \cdot (-3)$ should not mean that 6 does not factorise uniquely in $\mathbb{Z}$.

Definition 4.1 *Unique factorisation domain*

An integral domain R is a **unique factorisation domain** (UFD) if

(a) every non-zero non-unit element of R is expressible as a product of irreducible elements of R, and

(b) if $a \in R$ can be written as a product of irreducibles in two different ways,

$$a = p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_n,$$

then $m = n$ and the p_i, q_j pair off as associates.

Condition (a) of this definition tells us that a factorisation into irreducibles must *exist*, while part (b) ensures the *uniqueness*, up to associates. Our first example of a UFD is the ring of integers $\mathbb{Z}$: this, of course, is exactly what the Fundamental Theorem of Arithmetic tells us. On the other hand, we showed in Exercise 2.6 that the ring $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

We are now, at last, in a position to complete the story about the significance of prime and irreducible elements. The next theorem gives us a different characterisation of UFDs. As we have found in the examples we have looked at so far, irreducible and prime coincide precisely when elements factorise uniquely. Of course, we still need to be certain that we can always find *some* factorisation into irreducibles (condition (a) of Definition 4.1).

Theorem 4.2

Let R be an integral domain in which every non-zero non-unit element has a factorisation into irreducible elements. Then R is a UFD if, and only if, every irreducible element in R is prime.

Proof Suppose first that every irreducible element in R is prime. We will use a 'minimal counter-example' proof. Suppose that there exists $a \in R$ with two factorisations into irreducibles,

$$a = p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_n,$$

where the p_i and q_j do not pair off as associates. By swapping the two factorisations if necessary we can ensure $m \leq n$, and by the Well-Ordering Principle we choose such an a so that m is as small as possible. Note that if $m = 1$, then a itself is irreducible and so cannot be a product of more than one irreducible. This would imply that $a = p_1 = q_1$, which is a contradiction, and therefore $m > 1$.

This is related to mathematical induction. Supposing the statement is false, we take a smallest counter-example, and use the fact that anything smaller cannot be a counter-example to derive a contradiction.

Now, since p_m is prime and divides $a = q_1 q_2 \cdots q_n$, p_m must divide q_j for some j. This means that $q_j = u p_m$ for some $u \in R$, and since q_j is irreducible, u must be a unit. Now, however, if we write $b = p_1 p_2 \cdots p_{m-1}$, then

$$b p_m = a = q_1 q_2 \cdots q_n = q_1 \cdots q_{j-1} (u p_m) q_{j+1} \cdots q_n.$$

Cancelling p_m from either side (as we may since R is an integral domain), we have

$$b = p_1 \cdots p_{m-1} = q_1 q_2 \cdots q_{j-1}(u q_{j+1}) \cdots q_n.$$

We chose m to be as small as possible, and therefore b must have unique factorisation. This means that $m - 1 = n - 1$ (and so $m = n$), and the p_i $(1 \leq i \leq m-1)$ pair off with the remaining q_k as associates. However, we also paired off p_m with q_j, which means we have paired off all the p_i with the q_k, and this is a contradiction.

Conversely, suppose that R is a UFD. Consider an irreducible p for which $p \mid ab$ for some $a, b \in R$. That is, there is an $h \in R$ such that $ph = ab$. Then by unique factorisation, $a = p_1 \cdots p_n$, $b = q_1 \cdots q_m$ and $h = h_1 \cdots h_r$ where $p_1, \ldots, p_n, q_1, \ldots, q_m, h_1, \ldots, h_r$ are irreducible elements in R. This gives

$$ph_1 \cdots h_r = p_1 \cdots p_n q_1 \cdots q_m.$$

By the uniqueness of the factorisation, either p is an associate of p_i for some $i \in \{1, \ldots, n\}$ or p is an associate of q_j for some $j \in \{1, \ldots, m\}$. Hence either p divides a or p divides b. ■

Note that as well as providing a characterisation of UFDs, Theorem 4.2 also shows us that the terms 'irreducible' and 'prime' must coincide in a UFD. Although just a direct consequence of the above theorem and Definition 4.1, we record it here as a corollary.

Corollary 4.3

In a unique factorisation domain, every irreducible element is prime.

An important application of this corollary comes via its contrapositive: if a ring contains an irreducible element that is not prime, then it cannot be a UFD.

For example, consider again the ring $\mathbb{Z}[\sqrt{-5}]$. We know from Exercise 2.6 that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD since the element 6 can be written as a product of irreducibles in two different ways, $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$. However, by using Corollary 4.3 we could have deduced this at the end of Exercise 2.5: the element 2 is irreducible but not prime, and so $\mathbb{Z}[\sqrt{-5}]$ certainly cannot be a UFD. Of course, it is more informative to find an explicit example of non-unique factorisation; so the process in Exercise 2.6 is still a valuable technique to understand.

The next exercise asks you to prove the existence of least common multiples in UFDs. Before you can do so, we need to generalise the notion of least common multiple to integral domains.

Definition 4.4 *Least common multiple*

Let a and b be elements of an integral domain R. An element ℓ is said to be a **least common multiple** (**lcm**) of a and b if it satisfies the following:

(a) a divides ℓ and b divides ℓ

(b) for any $h \in R$ such that a divides h and b divides h, then ℓ also divides h.

Exercise 4.1

Prove that if R is a UFD, $\text{lcm}(a, b)$ exists for any non-zero $a, b \in R$.

4.2 UFDs and Euclidean domains

Since 'irreducible' and 'prime' mean the same thing in both Euclidean domains and UFDs, we might naturally wonder whether there is a relationship between these two concepts. In particular, we consider two questions: do all Euclidean domains have unique factorisation, and conversely, is every UFD a Euclidean domain? In light of Theorem 4.2, we can answer the first of these questions straight away.

Theorem 4.5

Every Euclidean domain is a unique factorisation domain.

Proof By Proposition 3.11 and Theorem 4.2, it suffices to show that each non-zero non-unit element of a Euclidean domain has some factorisation into irreducibles. Let R be a Euclidean ring with norm N, and let $a \in R$. We will again use a minimal counter-example proof. Let

$$S = \{x \in R : x \text{ is not zero and not a unit and cannot be written as a product of irreducibles}\}.$$

We want to show that S is empty. So, to derive a contradiction, let us suppose that it is non-empty. Using the Well-Ordering Principle, we choose $x \in S$ such that $N(x)$ is as small as possible.

Now since x cannot be written as a product of irreducibles, it cannot itself be irreducible. Hence there are elements $a, b \in R$, neither of which is a unit, with $ab = x$. Then, using Lemma 3.5(d),

$$N(a) < N(ab) = N(x) \text{ and } N(b) < N(ab) = N(x).$$

Hence neither a nor b belongs to S and so both can be written as products of irreducibles. Thus x can be written as a product of irreducibles, contradicting our assumption that S is non-empty. ■

In Worked Exercise 3.9, we proved that $\mathbb{Z}[\sqrt{-5}]$ was not a Euclidean domain with respect to any norm, because there are pairs of numbers that did not have a highest common factor. We now have another route to prove this, by using the contrapositive statement to Theorem 4.5: a ring that is not a UFD cannot be a Euclidean domain. Since we know that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD (either by Exercise 2.6 or by the comments following Corollary 4.3), we can immediately deduce that $\mathbb{Z}[\sqrt{-5}]$ is not a Euclidean domain.

Recall that, by Corollary 4.3, in order to show non-uniqueness of factorisation it suffices to find an element that is irreducible but not prime. To do this, we follow the process used in Exercise 2.5, using a multiplicative norm to prove irreducibility and indivisibility. The following exercise invites you to try this.

Exercise 4.2

Show that $\mathbb{Z}[\sqrt{-6}]$ is not a Euclidean domain.

Hint: using the norm $N(a+b\sqrt{-6}) = a^2+6b^2$, consider the element 2, and factorisations of the element 6.

The second question that we asked at the start of this subsection was whether every UFD is a Euclidean domain. The answer, perhaps surprisingly, is 'no'.

Example 4.6 *A UFD that is not a Euclidean domain*

In Exercise 3.3 we established that $\mathbb{Z}[x]$ is not a Euclidean domain, by demonstrating that there cannot exist integer polynomials $s(x)$ and $t(x)$ for which $2s(x)+xt(x) = \text{hcf}(2,x) = 1$.

However, we will prove shortly that $\mathbb{Z}[x]$ is a UFD.

Note that $\mathbb{Z}[x]$ fails to be a Euclidean domain, even though hcfs exist in it, because we cannot find suitable polynomials $s(x)$ and $t(x)$. In fact, in a unique factorisation domain, hcfs *must* exist.

Proposition 4.7

Let R be a unique factorisation domain and $a, b \in R$, not both of which are zero. Then $\text{hcf}(a,b)$ exists in R.

Proof If one of a, b is zero, then the non-zero one will be the highest common factor. If both are non-zero, then a and b factorise uniquely into a product of irreducibles, and so the highest common factor is the product of irreducible factors that occur in both factorisations. ■

On the other hand, rings such as $\mathbb{Z}[\sqrt{-5}]$ do not have unique factorisation, and in Worked Exercise 3.7 we used this to find a pair of elements that does not have a highest common factor.

We now turn to proving that $\mathbb{Z}[x]$ is a UFD. To do this, we will use the characterisation of UFDs given by Theorem 4.2, which involves two steps: the first is to show that every polynomial in $\mathbb{Z}[x]$ can be factorised into a product of irreducibles, and the second is that every irreducible polynomial in $\mathbb{Z}[x]$ is prime. In Chapter 17, Book E, we will use the same approach to prove a more general result, namely that $R[x]$ is a UFD whenever R is.

You may find it helpful to re-read Subsections 4.2 and 4.3 of Chapter 11 before continuing.

As we work through the proof, we will need to recall several definitions and results from Chapter 11. Recall that for a polynomial $f \in \mathbb{Z}[x]$, the **content**, c_f, of f is the highest common factor of all of its coefficients, and that if $c_f = 1$ then we say that f is **primitive**.

Lemma 4.8

Every polynomial in $\mathbb{Z}[x]$ has a factorisation into irreducible polynomials in $\mathbb{Z}[x]$.

Proof By the Fundamental Theorem of Arithmetic, the content c_f of a polynomial $f \in \mathbb{Z}[x]$ has a unique factorisation $c_f = p_1 p_2 \cdots p_s$ in $\mathbb{Z}$. Since each p_i is irreducible in $\mathbb{Z}$, it is also irreducible in $\mathbb{Z}[x]$, and so we can write $f = c_f f^*$ where f^* is primitive, and c_f has a unique factorisation into irreducibles. Consequently, it suffices to prove the lemma only for primitive polynomials in $\mathbb{Z}[x]$. Define

Remember that the irreducible elements of $\mathbb{Z}$ are the prime numbers.

$$S = \{f \in \mathbb{Z}[x] : f(x) \text{ is primitive, non-zero, is not } \pm 1 \text{ and cannot be written as a product of irreducibles}\}.$$

We need to show that S is empty. Suppose to the contrary that S is non-empty, and by the Well-Ordering Principle take a minimal counter-example. That is, take $f(x) \in S$ such that $\deg(f)$ is as small as possible.

Notice the similarity between this proof and the proof of Theorem 4.5.

Since f cannot be written as a product of irreducibles, it cannot itself be irreducible, and so it must have a proper factorisation $f = ab$ for some polynomials $a, b \in \mathbb{Z}[x]$. Since f is primitive, Lemma 4.12 of Chapter 11 tells us that both a and b are also primitive, and so in order for $f = ab$ to be a proper factorisation, we must have $\deg(a) \geq 1$ and $\deg(b) \geq 1$.

From $\deg(f) = \deg(ab) = \deg(a) + \deg(b)$ we know that $\deg(a) < \deg(f)$ and $\deg(b) < \deg(f)$. Since f was a polynomial of minimal degree in S, each of a and b is either irreducible, or can be written as a product of irreducibles. In every case, however, we reach a contradiction: $f = ab$ can be written as a product of irreducibles, even though $f \in S$. Consequently S is empty, and all primitive polynomials in $\mathbb{Z}[x]$ can be written as a product of irreducibles. ■

The second step needed to complete the proof that $\mathbb{Z}[x]$ is a UFD is to show that every irreducible element of $\mathbb{Z}[x]$ is prime.

Proposition 4.9

Let f be an irreducible polynomial in $\mathbb{Z}[x]$. Then f is prime in $\mathbb{Z}[x]$.

We divide the proof into two cases. The first is where we are given a polynomial $f \in \mathbb{Z}[x]$ for which $\deg(f) = 0$. In this case f is an integer, and since irreducible and prime mean the same thing in $\mathbb{Z}$, f is prime in $\mathbb{Z}$. The following lemma completes this case.

Lemma 4.10

Let $p \in \mathbb{Z}$. If p is prime in $\mathbb{Z}$, then it is prime in $\mathbb{Z}[x]$.

Proof Following a minimal counter-example argument, suppose that p is prime in $\mathbb{Z}$ but not in $\mathbb{Z}[x]$. Then there are polynomials $f, g \in \mathbb{Z}[x]$ such that p divides fg, but p divides neither f nor g. If $\deg(f) = m$ and $\deg(g) = n$, then $\deg(fg) = m + n$, and by the Well-Ordering Principle we can choose f and g so that $m + n$ is as small as possible.

Now from the fact that p divides fg it follows that p must divide every coefficient of fg. Denote the leading coefficients of f and g by a_m and b_n, respectively, so that the leading coefficient of fg is $a_m b_n$. Since p is prime in $\mathbb{Z}$, p must divide at least one of a_m or b_n, and without loss of generality suppose that p divides a_m. Note that $m > 0$ because otherwise p would divide f. Next, p divides

$$f(x)g(x) - a_m x^m g(x) = (f(x) - a_m x^m) \cdot g(x),$$

and $\deg(f(x) - a_m x^m) < m$. So, by our assumption that $m + n$ was as small as possible, p must divide $f(x) - a_m x^m$ since it does not divide g. However, p divides a_m, and so it divides $f(x) - a_m x^m + a_m x^m = f(x)$, which is a contradiction. ■

The other case, where $f \in \mathbb{Z}[x]$ and $\deg(f) \geq 1$, requires a different approach. The key step is to consider a polynomial in $\mathbb{Z}[x]$ as a polynomial in $\mathbb{Q}[x]$, and use the fact that irreducible and prime mean the same thing in $\mathbb{Q}[x]$ (since it is a Euclidean domain – see Example 3.2(c) of this chapter). This approach now enables us to give the formal proof of Proposition 4.9.

Proof of Proposition 4.9 Following the pattern laid out in the discussion above, we identify two cases. First, if $\deg(f) = 0$ then $f \in \mathbb{Z}$ is irreducible in $\mathbb{Z}$, and hence (since $\mathbb{Z}$ is a Euclidean domain) f is also prime in $\mathbb{Z}$. By Lemma 4.10, f is prime in $\mathbb{Z}[x]$.

Now suppose that $\deg(f) \geq 1$. Take $g, h \in \mathbb{Z}[x]$ for which f divides gh. To prove that f is prime in $\mathbb{Z}[x]$, we need to show that f divides one of g or h.

Since f is irreducible in $\mathbb{Z}[x]$ and $\deg(f) \geq 1$, f is primitive (otherwise $f = c_f f^*$ is a proper factorisation). Therefore, by Gauss's Lemma (Theorem 4.13 of Chapter 11), f is irreducible in $\mathbb{Q}[x]$. Since $\mathbb{Q}[x]$ is a

Euclidean domain, f is prime in $\mathbb{Q}[x]$, which means that in $\mathbb{Q}[x]$ either f divides g or f divides h, and without loss of generality we assume that f divides g. Therefore there exists $s \in \mathbb{Q}[x]$ such that $fs = g$. We want to show that the coefficients of $s(x)$ are all integers, so that f divides g in $\mathbb{Z}[x]$. Writing $s(x) = \frac{p}{q}t(x)$ where $t(x) \in \mathbb{Z}[x]$ is primitive and $\frac{p}{q} \in \mathbb{Q}$ is written in its lowest terms, we have, in $\mathbb{Q}[x]$,

The comments before Lemma 4.6 of Chapter 11 explain why we can write $s(x) = \frac{p}{q}t(x)$.

$$f(x)\left(\frac{p}{q}t(x)\right) = g(x).$$

Multiplying through by q, we obtain an equation in $\mathbb{Z}[x]$,

$$f(x)p\,t(x) = q\,g(x).$$

Now consider the content of each side of this equation. On the left, f and t are both primitive, and so the content is p. On the right, the content is qc_g, where c_g is the content of g. The two sides must be equal, and so $p = qc_g$, but this implies that $\frac{p}{q} = c_g$ is an integer. Therefore $s(x) = c_g t(x)$, and so $s \in \mathbb{Z}[x]$ and hence f divides g in $\mathbb{Z}[x]$. ■

Combining Lemma 4.8 with Proposition 4.9, we conclude the following.

Theorem 4.11

The ring $\mathbb{Z}[x]$ is a unique factorisation domain.

Exercise 4.3

(a) Prove that if f is a prime polynomial in $\mathbb{Z}[x]$ and $\deg(f) \geq 1$, then f is prime in $\mathbb{Q}[x]$.

(b) What happens if $f \in \mathbb{Z}[x]$ is prime, but $\deg(f) = 0$: is f still prime in $\mathbb{Q}[x]$?

4.3 Sums of two squares, revisited

In Subsection 3.3, we established that a prime number p can be expressed as the sum of two squares if, and only if, $p = 2$, or p is of the form $4k + 1$ for some $k \in \mathbb{N}$. We also found, in Worked Exercise 3.12 and Exercise 3.4, two examples of integers that can be expressed as a sum of squares in more than one way. This is not true of primes. No prime of the form $4k + 3$ can be expressed as a sum of two squares, whilst each prime of the form $4k + 1$ can be expressed as a sum of two squares *in a unique way*. The uniqueness assumes, as we have been doing all along, that by squares we mean squares of non-negative integers, and also that the sum of squares $x^2 + y^2$ is regarded as the same expression as $y^2 + x^2$.

In Theorem 3.13 we used the ring of Gaussian integers $\mathbb{Z}[i]$, and the fact that it is a Euclidean domain, to prove the existence of $x, y \in \mathbb{Z}$ such that $p = x^2 + y^2$ whenever p is of the form $4k + 1$. By Theorem 4.5 we now know that $\mathbb{Z}[i]$ is in fact a UFD, and this will enable us to prove the uniqueness of this expression.

Theorem 4.12 *Uniqueness of representation*

The expression of a prime p of the form $4k + 1$ as a sum of two squares is unique except for the order of the two summands.

Proof Suppose that $p = a^2 + b^2 = c^2 + d^2$, where $a, b, c, d \in \mathbb{Z}$. We must show that $a = \pm c$, $b = \pm d$ or $a = \pm d$, $b = \pm c$ so that the expression for p as a sum of two squares is unique (up to associates).

In $\mathbb{Z}[i]$, we have $p = (a + bi)(a - bi) = (c + di)(c - di)$. We claim that $a + bi$ and $a - bi$ are irreducible in $\mathbb{Z}[i]$. With the usual norm, we have $N(a + bi) = |a + bi|^2 = a^2 + b^2 = p$. Since p is prime in $\mathbb{Z}$, the only divisors of p in $\mathbb{Z}$ are 1 and p. Using the fact that N is multiplicative, for any factorisation of $a + bi = w \cdot z$ with $w, z \in \mathbb{Z}[i]$, either $N(w) = 1$ or $N(z) = 1$, and so one of w or z is a unit. Therefore $a + bi$ is irreducible. By using similar reasoning, it can also be shown that $a - bi$, $c + di$ and $c - di$ are irreducible in $\mathbb{Z}[i]$.

Now, since $\mathbb{Z}[i]$ is a UFD, the factorisation of $p = (a + bi)(a - bi)$ into irreducibles is unique up to associates, and thus either $(a + bi) = u(c + di)$ or $(a + bi) = v(c - di)$ where $u, v \in \{1, -1, i, -i\}$ are units. In either case, we compare real and imaginary coefficients to conclude that either $a = \pm c, b = \pm d$ or $a = \pm d, b = \pm c$. ■

We now know exactly which primes can be expressed as a sum of two squares, namely 2 and any prime of the form $4k + 1$, and we also know that any product of numbers expressible as a sum of two squares is itself so expressible. Hence any number that has no prime divisor of the form $4k + 3$ is certainly expressible as a sum of two squares. But are such numbers the only ones? A little experimentation reveals that they are not. For example, $18 = 3^2 + 3^2$ is a sum of two squares that does not adhere to the above prescription because it has a prime divisor of the form $4k + 3$, namely 3 itself. The essential point with 18 is that it is immaterial that 3 cannot be written as a sum of two squares, because 3^2 divides 18 and

$$18 = 2 \times 3^2 = (1^2 + 1^2)3^2 = 3^2 + 3^2.$$

In general, the identity

$$c^2(a^2 + b^2) = (ca)^2 + (cb)^2$$

shows that we can multiply any sum of squares by *any* square and retain a sum of squares. We can build on this observation in the following way. Suppose, for example, that we want to find one way of expressing

$2^7 \times 3^4 \times 5 \times 13^3$ as a sum of squares. By first isolating as large a square term as possible, we can write

$$2^7 \times 3^4 \times 5 \times 13^3 = (2^3 \times 3^2 \times 13)^2 \times (2 \times 5 \times 13)$$

Remember that a square-free integer is one that is not divisible by the square of any prime. Every positive integer is square, square-free or can be expressed as a square multiplied by a square-free integer.

and the task will be completed when we express the square-free part, $2 \times 5 \times 13$, as a sum of two squares.

$$\begin{aligned}(2^3 \times 3^2 \times 13)^2 \times 2 \times 5 \times 13 &= (2^3 \times 3^2 \times 13)^2 \times (1^2+1^2)(1^2+2^2)(2^2+3^2)\\ &= (2^3 \times 3^2 \times 13)^2 \times (1^2+1^2)(8^2+1^2)\\ &= (2^3 \times 3^2 \times 13)^2 \times (9^2+7^2)\\ &= (2^3 \times 3^4 \times 13)^2 + (2^3 \times 3^2 \times 7 \times 13)^2\end{aligned}$$

We make use of this idea in our proof of the following result, which gives a complete classification of which integers are sums of two squares.

Proposition 4.13 *Sums of two squares*

A positive integer n can be expressed as a sum of two squares if, and only if, each of its prime divisors of the form $4k+3$, if any, occurs to an even power.

Proof By pulling out the largest square divisor of n we can write $n = m^2 r$, where r is square-free. The theorem asserts that n is expressible as a sum of two squares if, and only if, r is not divisible by any prime of the form $4k+3$.

We allow the possibility $r = 1$.

To prove the 'if, and only if' assertion we have to establish the implications both ways; so we break the proof into two parts.

(a) Suppose that r has no prime divisor of the form $4k+3$. If $r = 1$ then $n = m^2 + 0^2$ and there is nothing to prove. If $r > 1$ then r is a product of one or more primes, each of which is either 2 or of the form $4k+1$. We have seen that such a product r can be expressed as a sum of two squares, and so $n = m^2(a^2+b^2) = (ma)^2 + (mb)^2$.

(b) Suppose that n can be expressed as a sum of two squares, say

$$n = m^2 r = a^2 + b^2.$$

First, any common divisor of a and b may be cancelled as follows. If $\mathrm{hcf}(a,b) = d$ then we can write $a = a_1 d$, $b = b_1 d$, where $\mathrm{hcf}(a_1, b_1) = 1$, and

$$m^2 r = d^2(a_1^2 + b_1^2).$$

Any prime that divides d occurs with exponent 2, or more, on the right-hand side of this equation and so must divide m.

As r is square-free, d divides m. Writing $m_1 = \frac{m}{d}$ gives

$$m_1^2 r = a_1^2 + b_1^2.$$

Our task is to show that r does not have a prime divisor of the form $4k+3$. Aiming for a contradiction, suppose that the prime $p = 4k+3$ divides r. Then

$$a_1^2 + b_1^2 \equiv 0 \,(\mathrm{mod}\ p); \quad \text{that is, } a_1^2 \equiv -b_1^2 \,(\mathrm{mod}\ p).$$

Now if p divides a_1 we would have p divides b_1, thereby contradicting $\text{hcf}(a_1, b_1) = 1$. So $\text{hcf}(a_1, p) = \text{hcf}(b_1, p) = 1$ and FLT can be applied, giving

$$a_1^{p-1} \equiv b_1^{p-1} \equiv 1 \,(\text{mod } p).$$

Putting $p = 4k + 3$,

$$\begin{aligned} 1 \equiv a_1^{p-1} &= a_1^{4k+2} = (a_1^2)^{2k+1} \equiv (-b_1^2)^{2k+1} \\ &\equiv (-1)^{2k+1}(b_1^2)^{2k+1} \equiv (-1)b_1^{p-1} \equiv -1 \,(\text{mod } p). \end{aligned}$$

This cannot possibly be true for an odd prime p and so we have a contradiction, and the result follows. ■

Exercise 4.4

(a) Which integers in the range 2000 to 2010 inclusive can be expressed as a sum of two squares? For any that can be so expressed, find one such representation.

(b) Use Proposition 4.13 to show that no positive integer that is congruent to either 3 or 6 modulo 9 can be expressed as a sum of two squares. For each of the other seven congruence classes modulo 9, find the smallest positive integer that can be represented as a sum of two squares and the smallest that cannot.

4.4 More Diophantine equations

We finish with two more applications of UFDs to solving Diophantine equations. First, using $\mathbb{Z}[\sqrt{-2}]$, we have the following result.

Proposition 4.14

The equation $a^2 + 2 = b^3$ has exactly one positive integer solution, namely $a = 5, b = 3$.

Proof There are two parts to this proof. First, we show

$$\text{hcf}(a + \sqrt{-2}, a - \sqrt{-2}) = 1;$$

so both $a - \sqrt{-2}$ and $a + \sqrt{-2}$ are cubes in $\mathbb{Z}[\sqrt{-2}]$. Next, we use the fact that

$$a - \sqrt{-2} = (c + d\sqrt{-2})^3$$

to conclude that $a = 5$ and $b = 3$.

We will work in the Euclidean domain $\mathbb{Z}[\sqrt{-2}]$. Here, $a^2 + 2 = (a + \sqrt{-2})(a - \sqrt{-2})$. We first claim that $\text{hcf}(a + \sqrt{-2}, a - \sqrt{-2}) = 1$.

Heading for a contradiction, suppose that p is an irreducible (and hence prime) element of $\mathbb{Z}[\sqrt{-2}]$ dividing $a+\sqrt{-2}$ and $a-\sqrt{-2}$. Then p divides $(a+\sqrt{-2})-(a-\sqrt{-2}) = 2\sqrt{-2} = (\sqrt{-2})^3$. By considering norms, $\sqrt{-2}$ is irreducible in $\mathbb{Z}[\sqrt{-2}]$, and since p is also irreducible, unique factorisation means that $p = \sqrt{-2}$ (or $p = -\sqrt{-2}$).

Now, $p = \sqrt{-2}$ divides $a+\sqrt{-2}$, and so $N(p) = 2$ divides $N(a+\sqrt{-2}) = a^2+2$. In other words, a^2+2 is even, and therefore a must be even. This in turn implies that $a^2+2 = 4k+2$ for some integer k. Next, from $a^2+2 = b^3$, we find that b^3 must be even. The only way for this to happen is if b is even, and then b^3 is of the form 8ℓ for some integer ℓ. Thus $4k+2 = 8\ell$. Dividing by 2 we find $2k+1 = 4\ell$, but the left-hand side is odd and the right-hand side is even, which is impossible. Thus there can be no irreducible element p dividing both $a+\sqrt{-2}$ and $a-\sqrt{-2}$, and hence $\text{hcf}(a+\sqrt{-2}, a-\sqrt{-2}) = 1$.

Finally, given $(a+\sqrt{-2})(a-\sqrt{-2}) = b^3$, and since $a+\sqrt{-2}$ and $a-\sqrt{-2}$ have a highest common factor of 1 in $\mathbb{Z}[\sqrt{-2}]$, they must both themselves be cubes. Then we have $a-\sqrt{-2} = (c+d\sqrt{-2})^3$ for some $c, d \in \mathbb{Z}$, which gives us

$$a-\sqrt{-2} = (c^3-6cd^2)+(3c^2d-2d^3)\sqrt{-2}.$$

Comparing coefficients, we have

$$\begin{aligned} a &= c^3-6cd^2 \\ -1 &= 3c^2d-2d^3 = d(3c^2-2d^2). \end{aligned}$$

The second equation implies $d = \pm 1$, which leaves us with $3c^2-2 = \mp 1$. If $d=-1$, then $3c^2 = 1$ has no solution in the integers, and so we must have $d=1$. Therefore $3c^2 = 3$, and so $c = \pm 1$.

Now, using the equation $a = c^3-6cd^2$, we have $a = (\pm 1)^3 \mp 6\cdot 1\cdot 1 = \mp 5$. Taking the positive solution, we get $a=5$, and then $b^3 = 27$ implies $b=3$, as required. ■

We finish with an exercise for you to try, which will lead you through the steps to prove that $a^2+1 = b^3$ has no non-trivial solutions in $\mathbb{Z}$.

Exercise 4.5

Let $a, b \in \mathbb{Z}$.

(a) Prove that if $a^2+1 = b^3$, then b must be odd, and hence a is even.

(b) Show that $1+i$, $1-i$ are irreducible elements in $\mathbb{Z}[i]$.

(c) Using parts (a) and (b), show that $\text{hcf}(a+i, a-i) = 1$ in $\mathbb{Z}[i]$.

Hint: use the fact that $2 = (1+i)(1-i)$.

(d) Hence conclude that the only solution to $a^2+1 = b^3$ in $\mathbb{Z}$ is $a=0$ and $b=1$.

Learning outcomes

We list those topics on which we may set assessment questions to test your understanding of the chapter.

After working through this chapter you should:

(a) Understand the term Pythagorean triple, recognise when a triple is primitive, and understand the statement and proof classifying the primitive Pythagorean triples.

(b) Understand the method of infinite descent, and be able to apply it to prove that certain Diophantine equations have no solutions.

(c) Know the meaning of the terms associate, factor, irreducible and prime as applied to elements of an arbitrary ring.

(d) Know what is meant by an integral domain, and understand why in an integral domain every prime element is irreducible, but not every irreducible element is prime.

(e) Recall the meaning of norm, multiplicative norm and Euclidean norm, apply these concepts to prove the irreducibility of ring elements, and use this to establish that in certain rings factorisation is not unique.

(f) Know the meaning of Euclidean domain, recall the approach to show that $\mathbb{Z}[i]$ is a Euclidean domain by finding the nearest point to a ring element, and be able to apply this approach to other rings.

(g) Know the definition of highest common factor of two elements of a ring, recall the approach to prove that highest common factors exist in Euclidean domains, and thus be able to show that certain rings are not Euclidean domains.

(h) Understand the steps required to prove that irreducible and prime elements coincide in a Euclidean domain.

(i) Understand the role Euclidean domains play in classifying when a prime number can be expressed as a sum of two squares, and use the method of exhaustion to express integers as a sum of two squares.

(j) Recall the definition of unique factorisation domain, the proof that every irreducible element is prime in a UFD, and that every Euclidean domain is a UFD.

(k) Be able to show that certain rings are not Euclidean domains by proving that they are not unique factorisation domains.

(l) Understand the key steps in the proof that $\mathbb{Z}[x]$ is a unique factorisation domain.

(m) Know why the fact that $\mathbb{Z}[i]$ is a UFD implies that primes of the form $4k+1$ are uniquely expressible as a sum of two squares except for the order of the two summands, and apply methods to show when a positive integer can or cannot be expressed as a sum of two squares.

(n) Be able to apply the theory of UFDs to compute all solutions (or show that none exist) to certain Diophantine equations.

Solutions to exercises

Solution to Exercise 1.1

(a) $(10, 8, 6)$ is not a Pythagorean triple; the hypotenuse must be the largest of the three values and, by convention, must be the last member of the triple.

(b) $(12, 9, 15)$ is a Pythagorean triple, but it is not primitive since it is the $(4, 3, 5)$ triple scaled by a multiple of 3.

(c) $(6, 7, 8)$ is not a Pythagorean triple: $6^2 + 7^2 \neq 8^2$.

(d) $(5, 12, 13)$ is a primitive Pythagorean triple: $5^2 + 12^2 = 169 = 13^2$ and the three integers are pairwise coprime.

(e) $(24, 33, 41)$ is not a Pythagorean triple since 24 and 33 have a common factor, namely 3, that does not divide the third number.

Solution to Exercise 1.2

(a) We seek coprime natural numbers m and n, of opposite parity and with $m > n$, such that $2mn = 72$. There are two such pairs:

$m = 36,\ n = 1,$ giving the triple $(72, 1295, 1297)$;

$m = 9,\ n = 4,$ giving the triple $(72, 65, 97)$.

The pairs $m = 18$, $n = 2$ and $m = 12$, $n = 3$ are excluded because they are not coprime.

(b) One possible strategy is to look at each divisor of 30 in turn and to determine all primitive triples containing that divisor. But notice that, since 30 is not a multiple of 4, no even divisor of 30 (including 30 itself) can occur in a primitive Pythagorean triple. So we need look only at the odd divisors of 30. Any primitive triple containing one of 3, 5 or 15 may be scaled appropriately to give a Pythagorean triple containing 30.

Note that the even member in a primitive triple is divisible by 4, since mn is even and $x = 2mn$.

We therefore seek all pairs of coprime natural numbers m and n of opposite parity and with $m > n$ such that either $m^2 + n^2$ or $m^2 - n^2$ is equal to one of 3, 5 or 15. It turns out that there are just five such pairs, as presented below.

Divisor	m	n	Primitive triple	Scaling	Triple with 30
15	4	1	$(8, 15, 17)$	2	$(16, 30, 34)$
15	8	7	$(112, 15, 113)$	2	$(224, 30, 226)$
5	3	2	$(12, 5, 13)$	6	$(72, 30, 78)$
5	2	1	$(4, 3, 5)$	6	$(24, 18, 30)$
3	2	1	$(4, 3, 5)$	10	$(40, 30, 50)$

As $m^2 - n^2 = (m - n)(m + n)$ we need only consider m and n for which $m + n \leq 15$.

(c) Suppose that $st = u^2$, where s and t are coprime. Let s and t have prime decompositions $s = p_1^{k_1} p_2^{k_2} \cdots p_i^{k_i}$, $t = p_{i+1}^{k_{i+1}} p_{i+2}^{k_{i+2}} \cdots p_j^{k_j}$. As $\mathrm{hcf}(s,t) = 1$, the listed primes are all distinct and so st has prime decomposition

$$st = p_1^{k_1} \cdots p_i^{k_i} p_{i+1}^{k_{i+1}} \cdots p_j^{k_j}.$$

The primes are not necessarily listed in ascending order here.

As st is a square, each exponent in this expression is even. Hence each exponent in s, and each exponent in t, is even and so these two are squares as well.

Solution to Exercise 1.3

Suppose that $x_1^4 + 2y_1^4 = 4z_1^4$, where x_1, y_1 and z_1 are positive integers. As the prime 2 divides two of the terms it must divide the third, namely x_1^4. So 2 divides x_1 and we can write $x_1 = 2x_2$ for some positive integer x_2. Substituting for x_1 gives

$$16x_2^4 + 2y_1^4 = 4z_1^4; \quad \text{that is, } 8x_2^4 + y_1^4 = 2z_1^4.$$

Repeating the argument, we can write $y_1 = 2y_2$ for some positive integer y_2 and obtain

$$8x_2^4 + 16y_2^4 = 2z_1^4; \quad \text{that is, } 4x_2^4 + 8y_2^4 = z_1^4.$$

We can now write $z_1 = 2z_2$ for some positive integer z_2 to obtain

$$4x_2^4 + 8y_2^4 = 16z_2^4; \quad \text{that is, } x_2^4 + 2y_2^4 = 4z_2^4.$$

At this point we have reached a second positive solution, $x = x_2$, $y = y_2$, $z = z_2$ of the original equation with $z_2 < z_1$. The descent step is therefore complete and the required contradiction is established.

Thus $x^4 + 2y^4 = 4z^4$ has no positive solutions.

Solution to Exercise 1.4

(a) As any square is congruent modulo 4 to 0 (if it is even) or to 1 (if it is odd), it follows that $x^2 + y^2 + z^2 \equiv 0 \pmod 4$ can occur only when x, y and z are all even.

(b) Suppose that $x_1^2 + y_1^2 + z_1^2 = 2x_1y_1z_1$. If x_1, y_1 and z_1 are all odd then $x_1^2 + y_1^2 + z_1^2 \equiv 3 \pmod 4$ while the right-hand side, $2x_1y_1z_1$, is even. Hence at least one of x_1, y_1 and z_1 must be even. But then $x_1^2 + y_1^2 + z_1^2 = 2x_1y_1z_1 \equiv 0 \pmod 4$ and, as we have just seen, x_1, y_1 and z_1 must all be even.

Writing $x_1 = 2x_2$, $y_1 = 2y_2$ and $z_1 = 2z_2$ gives

$$(2x_2)^2 + (2y_2)^2 + (2z_2)^2 = 2(2x_2)(2y_2)(2z_2),$$

which simplifies to

$$x_2^2 + y_2^2 + z_2^2 = 4x_2y_2z_2.$$

As the right-hand side is congruent modulo 4 to 0, the first part of the question once again gives that x_2, y_2, and z_2 are even positive integers.

(c) Continuing from part (b), if we write $x_2 = 2x_3$, $y_2 = 2y_3$, $z_2 = 2z_3$ we obtain

$$x_3^2 + y_3^2 + z_3^2 = 8x_3y_3z_3$$

with x_3, y_3 and z_3 being even positive integers. We can continue forever in this way, halving the x, y and z values and yet retaining even positive integers, for from

$$x_n^2 + y_n^2 + z_n^2 = 2^n x_n y_n z_n$$

we note that x_n, y_n and z_n are even integers and

$$\left(\frac{x_n}{2}\right)^2 + \left(\frac{y_n}{2}\right)^2 + \left(\frac{z_n}{2}\right)^2 = 2^{n+1}\left(\frac{x_n}{2}\right)\left(\frac{y_n}{2}\right)\left(\frac{z_n}{2}\right).$$

The same argument now confirms that $\frac{x_n}{2}$, $\frac{y_n}{2}$ and $\frac{z_n}{2}$ are still even positive integers. As the sequence $z_1, z_2, z_3, \ldots$ of positive integers cannot decrease indefinitely, the method of infinite descent shows that the supposed positive solution of the original equation cannot exist.

Solution to Exercise 1.5

From $x^2 + y^2 = z^2$ and $xy = 2n^2$ we have

$$(x+y)^2 = x^2 + y^2 + 2xy = z^2 + (2n)^2$$

and

$$(x-y)^2 = x^2 + y^2 - 2xy = z^2 - (2n)^2.$$

Multiplying these two equations together gives

$$(x^2 - y^2)^2 = z^4 - (2n)^4.$$

Note that $x \neq y$ in any Pythagorean triple and that $x^2 - y^2$ can be replaced by $y^2 - x^2$ if the former is negative.

By Corollary 1.10 no positive solution can exist. We conclude that no Pythagorean triangle can have an area that is a square.

Solution to Exercise 2.1

(a) If a, b satisfy $a^2 - db^2 = 1$, then in $\mathbb{Z}[\sqrt{d}]$ we have $(a + b\sqrt{d})(a - b\sqrt{d}) = a^2 - db^2 = 1$, and so $a + b\sqrt{d}$ is a unit.

(b) Since a, b are both greater than 0, the unit $u = a + b\sqrt{d}$ of $\mathbb{Z}[\sqrt{d}]$ is a real number that is strictly greater than 1. By Lemma 2.2, u^k is a unit of $\mathbb{Z}[\sqrt{d}]$ for all $k \geq 1$, and since $u > 1$ these are all distinct. Hence $\mathbb{Z}[\sqrt{d}]$ has infinitely many units when $d > 1$ is square-free.

Solution to Exercise 2.2

(a) (i) Using the multiplication table for $\mathbb{Z}_6$ given in Example 2.7, we find that the units are 1 and 5, since $5 \cdot 5 \equiv 1 \pmod 6$.

The non-zero non-units are therefore 2, 3 and 4. None is irreducible since $2 = 2 \cdot 4$, $3 = 3 \cdot 3$ and $4 = 2 \cdot 2$, and so there are no irreducible elements in $\mathbb{Z}_6$. However, all of 2, 3 and 4 are prime. Example 2.7 shows that 2 is prime. Next, 3 divides 0 and 3, but every product $a \cdot b \equiv 0$ or $3 \pmod 6$ requires that one of a or b is 0 or 3. Finally, we note that 4 divides 0, 2 and 4, but whenever $a \cdot b \equiv 0, 2$ or $4 \pmod 6$ we find that at least one of a or b is 0, 2 or 4.

(ii) The multiplication table for $\mathbb{Z}_8 \setminus \{0\}$ is:

	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	2	4	6	0	2	4	6
3	3	6	1	4	7	2	5
4	4	0	4	0	4	0	4
5	5	2	7	4	1	6	3
6	6	4	2	0	6	4	2
7	7	6	5	4	3	2	1

The units are $1, 3, 5$ and 7, since $3 \cdot 3 \equiv 1$, $5 \cdot 5 \equiv 1$ and $7 \cdot 7 \equiv 1 \pmod 8$.

The non-zero non-units are therefore 2, 4 and 6. The irreducible elements are 2 and 6 since whenever $a \cdot b \equiv 2$ or $6 \pmod 8$, one of a or b is a unit. 4 is not irreducible since $4 = 2 \cdot 2$.

The primes in $\mathbb{Z}_8$ are also 2 and 6. First, 4 is not prime since $4 = 2 \cdot 2$, but reading along the 4th row we see that 4 does not divide 2. To prove that 2 and 6 are prime, observe that both 2 and 6 divide the elements 0, 2, 4 and 6. Whenever $a \cdot b \equiv 0$, 2, 4 or $6 \pmod 8$ then one of a or b is equal to one of 0, 2, 4 or 6.

(b) No, since $(4 + 3i) = (1 + 2i)(2 - i)$. Do not worry if you did not find this factorisation. The point of this question was to make you aware that deciding irreducibility is not very easy.

Notice that

$$\begin{aligned} |4 + 3i|^2 &= 4^2 + 3^2 = 25 \\ &= 5 \cdot 5 = (1 + 4) \cdot (4 + 1) \\ &= |1 + 2i|^2 \cdot |2 - i|^2, \end{aligned}$$

which suggests how you might go about finding such obscure factorisations.

Solution to Exercise 2.3

(a) Since $a \mid c$, there exists $b \in R$ such that $c = ba$. Now, $c + na = ba + na = (b + n)a$ proves that a divides $c + na$.

(b) Since $a \mid b$, there exists non-zero $c \in R$ such that $b = ac$. Similarly, $b \mid a$ implies there exists non-zero $d \in R$ such that $a = bd$. Now $a = bd = acd$ and Lemma 2.11 now shows that $cd = 1$. So c and d are units, and hence a and b are associates.

(c) Since $a \mid b$ and $b \mid c$, there exist non-zero $p, q \in R$ such that $b = ap$ and $c = bq$. Now $c = bq = (ap)q = a(pq)$ and therefore a divides c.

(d) Since $a \mid b$ and $a \mid c$, there exist non-zero $p, q \in R$ such that $b = ap$ and $c = aq$. Now $mb + nc = map + naq = amp + anq = a(mp + nq)$ and so a divides $mb + nc$.

(e) This follows since $0 = a \cdot 0$ for all $a \in R$.

(f) This follows from $n = u(u^{-1}n)$.

Solution to Exercise 2.4

Suppose $S \subseteq R$ is not an integral domain. Then there exist non-zero $a, b \in S$ such that $a \cdot b = 0_S$, where 0_S denotes the additive identity of the ring S.

Since S is a subring of R, we know that $0_S = 0_R$. Therefore $a \cdot b = 0_R$ in R, and since $a, b \neq 0_R$ this shows that R cannot be an integral domain if S is not an integral domain.

Solution to Exercise 2.5

(a) Since R is a subring of the integral domain $\mathbb{C}$, it follows by Exercise 2.4 that R is an integral domain.

Alternatively, notice that N is the complex modulus squared, which we know to be multiplicative.

(b) Let $r = a + b\sqrt{-5}, s = c + d\sqrt{-5} \in R \setminus \{0\}$. Then

$$\begin{aligned} N(rs) &= N((ac - 5bd) + (bc + ad)\sqrt{-5}) \\ &= (ac - 5bd)^2 + 5(bc + ad)^2 \\ &= a^2c^2 - 10abcd + 25b^2d^2 + 5b^2c^2 + 10abcd + 5a^2d^2 \\ &= a^2c^2 + 25b^2d^2 + 5b^2c^2 + 5a^2d^2 \\ &= (a^2 + 5b^2)(c^2 + 5d^2) = N(r)N(s). \end{aligned}$$

(c) The result follows immediately since if $b \neq 0$ then $b^2 \geq 1$ and so $a^2 + 5b^2 \geq 5$. If $N(r) < 5$ then $b = 0$, and so $r \in \mathbb{Z} \setminus \{0\}$. In fact, the only possibilities are $r = \pm 1, \pm 2$.

(d) Suppose $2 = rs$ for some $r, s \in R$. By part (b), $N(r)$ and $N(s)$ both divide $N(2) = 4$. Therefore, we have $N(r), N(s) < 5$, and so by part (c) r and s must be integers. However, 2 does not have any proper integer divisors, and so in R it is irreducible.

(e) If 2 divided $(1+\sqrt{-5})$ or $(1-\sqrt{-5})$ then, since N is multiplicative, $N(2)$ would divide $N(1+\sqrt{-5})$ or $N(1-\sqrt{-5})$. However, $N(2)=4$ does not divide $N(1+\sqrt{-5})=N(1-\sqrt{-5})=6$, and so 2 divides neither $(1+\sqrt{-5})$ nor $(1-\sqrt{-5})$.

This proves that 2 is not prime, since 2 divides $2\cdot 3=6$ $=(1+\sqrt{-5})(1-\sqrt{-5})$, but it divides neither $(1+\sqrt{-5})$ nor $(1-\sqrt{-5})$.

Solution to Exercise 2.6

(a) If $u=a+b\sqrt{-5}$ is a unit in R, then there is $u^{-1}\in R$ with $uu^{-1}=1$. This implies that $N(u)N(u^{-1})=N(1)=1$. Since $N(u),N(u^{-1})\geq 1$ we must have $N(u)=1$. Now $N(u)=N(a+b\sqrt{-5})=a^2+5b^2$, and the only solutions to $a^2+5b^2=1$ are $a=\pm 1, b=0$. Therefore, the only elements in R with $N(u)=1$ are $1,-1$.

Note that since N is defined only for non-zero elements of R, we have $N(a+b\sqrt{-5})>0$.

Finally, $1\cdot 1=1$ and $(-1)\cdot(-1)=1$ proves that both 1 and -1 are indeed units.

(b) First, we compute the norm: $N(3)=3^2+0=9$. If 3 is reducible, then $3=rs$ where $r,s\in R$ are not units. Since N is multiplicative, this means that $N(r)N(s)=N(rs)=N(3)=9$. Next, r and s are not units, and so $N(r)>1$ and $N(s)>1$ (using part (a)), and therefore $N(r)=N(s)=3$.

From Exercise 2.5(c), the only non-zero elements of R for which $N(r)<5$ are integers, with $r=\pm 1,\pm 2$. Now, $N(\pm 1)=1$ and $N(\pm 2)=4$, and so there are no elements $r\in R$ with $N(r)=3$. Therefore we cannot have $3=rs$, unless one of r or s is a unit.

(c) Suppose that $r,s\in R$ are not units, with $1+\sqrt{-5}=rs$. Then $N(rs)=N(1+\sqrt{-5})=6=N(r)N(s)$. Since $N(r)>1$ and $N(s)>1$, the norms of r and s must be 2 and 3, in some order. However, in part (b) we proved that there are no elements of norm 3, and so this is impossible. Thus $1+\sqrt{-5}$ is irreducible.

The same argument shows that $1-\sqrt{-5}$ is irreducible.

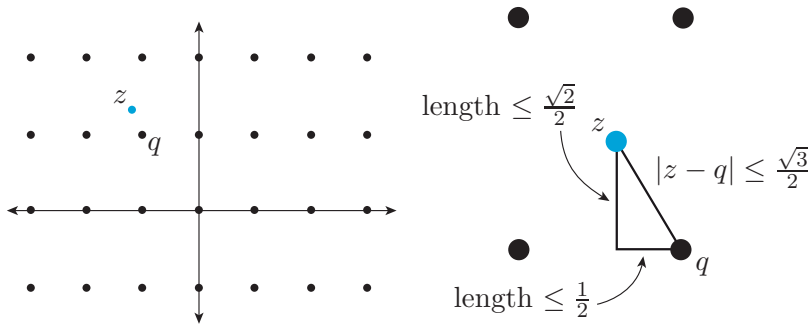
(d) $6=(1+\sqrt{-5})(1-\sqrt{-5})=2\cdot 3$, where all four factors are irreducible, and none is an associate of another since the only units are 1 and -1.

Solution to Exercise 3.1

(a) We follow the approach laid out for $\mathbb{Z}[i]$ in Worked Exercise 3.4. First, $\mathbb{Z}[\sqrt{-2}]$ is an integral domain since it is a subring of $\mathbb{C}$.

We now show that N is a Euclidean norm. For non-zero $a,b\in\mathbb{Z}[\sqrt{-2}]$, we have $N(ab)=N(a)N(b)\geq N(a)$ since $N(b)\geq 1$ unless $b=0$.

Taking $a,b\in\mathbb{Z}[\sqrt{-2}]$ and $b\neq 0$, to find q,r such that $a=bq+r$ we set q to be a point in $\mathbb{Z}[\sqrt{-2}]$ that is nearest to the complex number $z=\frac{a}{b}$.



The distance from z to q in this diagram is $|z-q|$. Since q is the nearest point of $\mathbb{Z}[\sqrt{-2}]$ to z, $|z-q|$ cannot be more than the distance from q to the centre of the rectangle bounded by points of $\mathbb{Z}[\sqrt{-2}]$ containing z. This rectangle has dimensions $1 \times \sqrt{2}$, and so the distance from a corner to the centre is $\frac{\sqrt{3}}{2}$. Therefore

$$|z-q| \leq \tfrac{\sqrt{3}}{2} < 1.$$

Now, set $r = a - bq$. Then $r \in \mathbb{Z}[\sqrt{-2}]$, and

$$N(r) = |r|^2 = |a-bq|^2 = |b|^2|z-q|^2 < |b|^2 = N(b),$$

which proves that N is a Euclidean norm on $\mathbb{Z}[\sqrt{-2}]$.

Note that 'failing to prove $\mathbb{Z}[\sqrt{-5}]$ is Euclidean' does not prove anything: it merely indicates why the earlier approach does not work.

(b) If we attempt to prove that $\mathbb{Z}[\sqrt{-5}]$ is a Euclidean domain by modifying the argument we used for part (a) of this exercise, we find that the points $0, \sqrt{-5}, 1, 1+\sqrt{-5}$ of $\mathbb{Z}[\sqrt{-5}]$ form the corners of a rectangle in the complex plane whose sides have lengths 1 and $\sqrt{5} \approx 2.236$. This implies that the centre of the rectangle is at a distance $\frac{\sqrt{6}}{2} \approx 1.223 > 1$ from any of the corners.

Thus, for $a, b \in \mathbb{Z}[\sqrt{-5}]$, with $b \neq 0$, it is possible that if $q \in \mathbb{Z}[\sqrt{-5}]$ is the nearest point to the complex number $z = \frac{a}{b}$, then we have $|z-q| > 1$. If this occurs, then setting $r = a - bq$ as before does not work, since $N(r) = N(a-bq) = |b|^2|z-q|^2 > N(b)$.

(c) Suppose there exist $q, r \in \mathbb{Z}[\sqrt{-5}]$ such that $1+\sqrt{-5} = 2q + r$ with either $r = 0$ or $N(r) < N(2)$. Because $N(2) = 2^2 = 4$ and $N(1+\sqrt{-5}) = 1+5 = 6$, we know that 2 does not divide $1+\sqrt{-5}$ and so $r \neq 0$. We therefore must have $N(r) < N(2) = 4$. From Exercise 2.5, the condition $N(r) < 4$ implies that $r = 1$ or -1.

Write $q = a + b\sqrt{-5}$ with $a, b \in \mathbb{Z}$. If $r = 1$ then $1+\sqrt{-5} = 2q+1$, and subtracting 1 from either side gives us

$$\sqrt{-5} = 2a + 2b\sqrt{-5}$$

from which we conclude $a = 0$ and $b = \frac{1}{2}$. However, b must be an integer, and so this is impossible.

Similarly, if $r = -1$, then $1+\sqrt{-5} = 2q - 1$, which tells us

$$2 + \sqrt{-5} = 2a + 2b\sqrt{-5}$$

and so $a = 1$, but again $b = \frac{1}{2}$, which is impossible. Therefore no such r can exist with $N(r) < N(2)$, and so N is not a Euclidean norm.

Solution to Exercise 3.2

(a) For non-zero $a+bi \in \mathbb{Z}[i]$, we will use the Euclidean norm $N(a+bi) = |a+bi|^2 = a^2+b^2$. From the fact that $N(1) = 1$, it follows that if $a+bi$ is a unit in $\mathbb{Z}[i]$ then it must satisfy $a^2+b^2 = 1$.

If $|a| > 1$ or $|b| > 1$, then $a^2+b^2 \geq 4$ and so there is no solution to $a^2+b^2 = 1$. We cannot have $a = b = 0$, or $|a| = |b| = 1$, and this leaves four cases: $a = \pm 1$ and $b = 0$, or $a = 0$ and $b = \pm 1$. Hence, the units in $\mathbb{Z}[i]$ are $\pm 1, \pm i$.

(b) Since $1 \in \mathbb{Z}[\sqrt{-2}]$ is a unit and $N(1) = 1^2 + 2 \cdot 0^2 = 1$, any unit $a+b\sqrt{-2}$ must satisfy $a^2+2b^2 = 1$. If $b \neq 0$ then $a^2+2b^2 > 1$, and so we must have $b = 0$. This leaves $a^2 = 1$, and so the only units of $\mathbb{Z}[\sqrt{-2}]$ are 1 and -1.

Solution to Exercise 3.3

(a) The only divisors of $2 \in \mathbb{Z}[x]$ are 1 and 2, and the divisors of $x \in \mathbb{Z}[x]$ are 1 and x. Therefore, 1 is a common factor, and is clearly maximal.

(b) If $\mathbb{Z}[x]$ is a Euclidean domain, then since $\text{hcf}(2, x) = 1$ there should exist polynomials $s, t \in \mathbb{Z}[x]$ such that

$$2s(x) + xt(x) = 1.$$

Notice that the constant term on the left-hand side of this equation must equal 1. However, since $x \cdot t(x)$ has constant term zero, the constant term of $2s(x)$ must equal 1, but this is impossible since all coefficients of $2s(x)$ must be divisible by 2. This tells us that $\mathbb{Z}[x]$ is not a Euclidean domain.

(c) By inspection, we may take $s(x) = \frac{1}{2}$ and $t(x) = 0$.

Solution to Exercise 3.4

$$\begin{aligned} 325 &= 5^2 \times 13 = (2^2+1^2)(2^2+1^2)(3^2+2^2) \\ &= (2^2+1^2)(8^2+1^2) = 17^2+6^2 \\ &= (1^2+2^2)(8^2+1^2) = 10^2+15^2 \\ &= (2^2+1^2)(2^2+3^2)(2^2+1^2) = (7^2+4^2)(2^2+1^2) = 18^2+1^2 \end{aligned}$$

Solution to Exercise 3.5

As $313 \equiv 1 \pmod 4$ we know that 313 can be expressed as a sum of two squares. Searching the sequence

$$313 - 1^2, \quad 313 - 2^2, \quad 313 - 3^2, \quad \ldots$$

for the first square, we find $313 - 12^2 = 13^2$. Therefore

$$5321 = 17 \times 313 = (1^2+4^2)(12^2+13^2) = 64^2+35^2.$$

In fact there is just one alternative solution, which can be reached by writing 313 as 13^2+12^2:

$$5321 = (1^2+4^2)(13^2+12^2) = 61^2+40^2.$$

The expression for a prime is unique, but it may not be so for a composite integer.

Solution to Exercise 4.1

For non-zero $a, b \in R$, since R is a UFD we can write a and b uniquely as products of irreducible elements in R. We fix irreducibles $p_1, p_2, \ldots, p_n$ and a unit $u \in R$ so that we can write

$$a = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}, \qquad b = u p_1^{b_1} p_2^{b_2} \cdots p_n^{b_n}$$

for some exponents $a_i, b_i \geq 0$ (note $a_i = 0$ or $b_i = 0$ is explicitly allowed). Note that the unit u simply recognises the fact that b could be an associate of the product $p_1^{b_1} p_2^{b_2} \cdots p_n^{b_n}$.

We claim $\ell = p_1^{\max(a_1,b_1)} p_2^{\max(a_2,b_2)} \cdots p_n^{\max(a_n,b_n)}$ is a least common multiple for a and b. Certainly a divides ℓ, since each irreducible power $p_i^{a_i}$ divides $p_i^{\max(a_i,b_i)}$, and similarly b divides ℓ.

Now suppose $h \in R$ is divisible by a and b. Then since the factorisation of h is unique, $p_i^{\max(a_i,b_i)}$ divides h for each $i = 1, 2, \ldots, n$, and so ℓ divides h, as required.

Solution to Exercise 4.2

We prove that 2 is irreducible in $\mathbb{Z}[\sqrt{-6}]$, but does not divide $\sqrt{-6}$, so that 2 is irreducible but not prime in $\mathbb{Z}[\sqrt{-6}]$, since $6 = 2 \cdot 3 = -(\sqrt{-6})^2$. Then, by Corollary 4.3 and the comments that follow it, we know that $\mathbb{Z}[\sqrt{-6}]$ is not a UFD, and hence not a Euclidean domain by Theorem 4.5.

Alternatively, you could show that the elements 2, 3 and $\sqrt{-6}$ are all irreducible in $\mathbb{Z}[\sqrt{-6}]$, and therefore $\mathbb{Z}[\sqrt{-6}]$ is not a UFD since $6 = 2 \cdot 3 = -(\sqrt{-6})^2$.

To prove that 2 is irreducible and does not divide $\sqrt{-6}$, we use the multiplicative norm $N(a + b\sqrt{-6}) = |a + b\sqrt{-6}|^2 = a^2 + 6b^2$. Since $N(2) = 4$ and N is multiplicative, any element dividing 2 must have norm dividing 4. If $N(a + b\sqrt{-6}) = 1$, then $a^2 + 6b^2 = 1$ implies that $a = \pm 1$ and $b = 0$, and so the only elements with norm 1 are the units $+1$ and -1. Therefore, if the element 2 is reducible then its two factors in $\mathbb{Z}[\sqrt{-6}]$ must both have norm equal to 2. However, there are no elements with norm 2 in $\mathbb{Z}[\sqrt{-6}]$, as there are no solutions to $a^2 + 6b^2 = 2$ in $\mathbb{Z}$: we must have $b = 0$ otherwise $a^2 + 6b^2 \geq 6$, but this leaves $a^2 = 2$, which has no solution in $\mathbb{Z}$. Thus there are no elements in $\mathbb{Z}[\sqrt{-6}]$ with norm 2, and therefore the element 2 is irreducible in $\mathbb{Z}[\sqrt{-6}]$.

Notice that we can't use results from Lemma 3.5 here, because that applies only to Euclidean domains.

Finally, since N is multiplicative and $N(2) = 4$ does not divide $N(\sqrt{-6}) = 6$, we conclude that 2 does not divide $\sqrt{-6}$, and is therefore irreducible but not prime. Hence $\mathbb{Z}[\sqrt{-6}]$ is not a UFD, and hence not a Euclidean domain.

Solution to Exercise 4.3

(a) Since f is prime in $\mathbb{Z}[x]$, it is irreducible in $\mathbb{Z}[x]$ by Theorem 2.12 of this chapter. Since $\deg(f) \geq 1$, f is primitive (otherwise $f = c_f f^*$ is a proper factorisation).

By Gauss's Lemma (Theorem 4.13 of Chapter 11), f is irreducible in $\mathbb{Q}[x]$. Finally, since $\mathbb{Q}[x]$ is a Euclidean domain, every irreducible element is prime, and so f is prime in $\mathbb{Q}[x]$.

(b) No – $\deg(f) = 0$ means that $f \in \mathbb{Z}$, and hence $f \in \mathbb{Q}$. However, $\mathbb{Q}$ is a field, and so f is a unit in $\mathbb{Q}$. All units of $\mathbb{Q}$ are units of $\mathbb{Q}[x]$, and so f cannot be prime in $\mathbb{Q}[x]$.

Solution to Exercise 4.4

(a)

$2000 = 2^4 \times 5^3$	$2000 = 40^2 + 20^2$
$2001 = 3 \times 23 \times 29$	not expressible since $3 \equiv 3 \pmod 4$
$2002 = 2 \times 7 \times 11 \times 13$	not expressible since $7 \equiv 3 \pmod 4$
2003 is a prime	not expressible since $2003 \equiv 3 \pmod 4$
$2004 = 2^2 \times 3 \times 167$	not expressible since $3 \equiv 3 \pmod 4$
$2005 = 5 \times 401$	$2005 = 41^2 + 18^2$
$2006 = 2 \times 17 \times 59$	not expressible since $59 \equiv 3 \pmod 4$
$2007 = 3^2 \times 223$	not expressible since $223 \equiv 3 \pmod 4$
$2008 = 2^3 \times 251$	not expressible since $251 \equiv 3 \pmod 4$
$2009 = 7^2 \times 41$	$2009 = 28^2 + 35^2$
$2010 = 2 \times 3 \times 5 \times 67$	not expressible since $3 \equiv 3 \pmod 4$

(b) If $n \equiv 3 \pmod 9$ then $n = 9k + 3 = 3(3k + 1)$ for some integer k. Now $\text{hcf}(3, 3k + 1) = 1$ and so, in the prime decomposition of n, 3 occurs with exponent 1. Proposition 4.13 confirms that n cannot be expressed as a sum of two squares.

Similarly, if $n = 9k + 6 = 3(3k + 2)$ then n is divisible by 3 but not by 3^2 and so it is not expressible as a sum of two squares.

Taking each of the other congruence classes modulo 9 in turn:

Residue class modulo 9	Smallest sum of two squares	Smallest not sum of two squares
0	$9 = 3^2 + 0^2$	$27 = 3^3$
1	$1 = 1^2 + 0^2$	19
2	$2 = 1^2 + 1^2$	11
4	$4 = 2^2 + 0^2$	$22 = 2 \times 11$
5	$5 = 2^2 + 1^2$	$14 = 2 \times 7$
7	$16 = 4^2 + 0^2$	7
8	$8 = 2^2 + 2^2$	$35 = 5 \times 7$

Each of the primes 3, 7, 11 and 19 is of the form $4k + 3$.

Solution to Exercise 4.5

(a) Suppose, heading for a contradiction, that b is even, and so $b = 2k$, which implies $b^3 = 8k$. From $a^2 + 1 = b^3 = 8k$, we have $a^2 = 8k - 1 = 4(2k - 1) + 3 \equiv 3 \pmod 4$. However, all square numbers are congruent to 0 or 1 modulo 4, and so there is no integer a for which $a^2 = 4(2k - 1) + 3$. Therefore, b must be odd, and so a must be even.

(b) This follows directly by taking norms: $N(1 + i) = |1 + i|^2 = 2$, and since N is multiplicative and 2 is prime in $\mathbb{Z}$, $1 + i$ can have no proper factors. Similarly, $N(1 - i) = 2$ gives the same conclusion.

(c) Suppose p is an irreducible element of $\mathbb{Z}[i]$ that divides $a + i$ and $a - i$. Then p divides $(a + i) - (a - i) = 2i$. Now, i is a unit, and we can write $2 = (1 + i)(1 - i)$. So if p is irreducible and divides $2i$, then by part (b) and the fact that $\mathbb{Z}[i]$ is a UFD we have $p = 1 + i$ or $p = 1 - i$.

If $p = 1 + i$, then since p divides $a + i$, we can find $c, d \in \mathbb{Z}$ such that $(1 + i)(c + di) = a + i$, from which we find $c - d = a$ and $c + d = 1$. Eliminating d, we get $2c = a + 1$, which tells us that a is odd, and this is impossible by part (a).

Similarly if $p = 1 - i$, then writing $a + i = (1 - i)(c + di)$ tells us that $a = c + d$ and $1 = d - c$. This gives $2c = a - 1$, again implying a is odd, which is impossible.

(d) Since $\text{hcf}(a + i, a - i) = 1$, if $(a + i)(a - i) = b^3$, then both $a + i$ and $a - i$ must be cubes in $\mathbb{Z}[i]$. Thus, we can find $c, d \in \mathbb{Z}$ such that $a + i = (c + di)^3$. Comparing real and imaginary parts, we have

$$a = c^3 - 3cd^2$$
$$1 = 3c^2d - d^3.$$

From this, we have $1 = d(3c^2 - d^2)$, which forces $d = \pm 1$. So then we must have $c = 0$ and therefore $a = 0$. Thus, the only solution to $a^2 + 1 = b^3$ is $a = 0$, $b = 1$.